CLC Genomics Workbench
ハンズオントレーニング
変異解析編

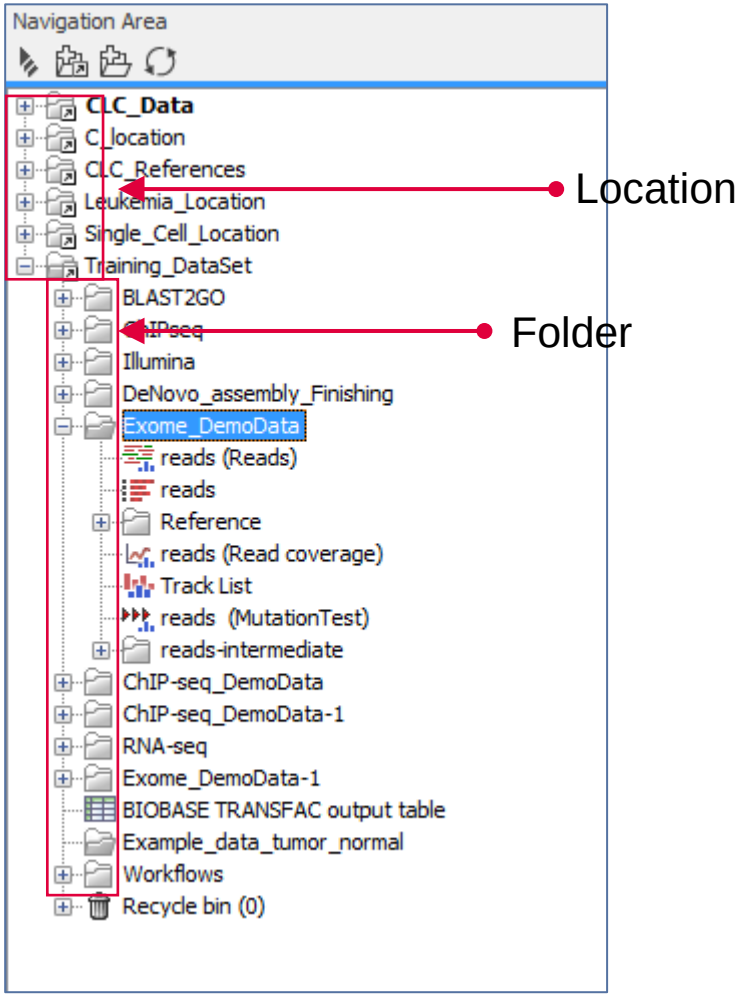
株式会社キアゲン
グローバルイフォマティクス ソリューション&サポート
アプライドアドバンストゲノミクス

Genomics Workbench で可能な解析

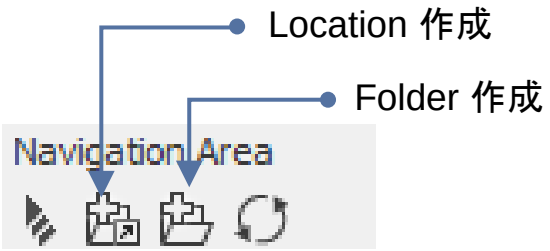
新規生物種	変異解析	ChIP-seq	RNA-seq	small RNA
インポート	インポート	インポート	インポート	インポート
Quality check	Quality check	Quality check	Quality check	タグの抽出
De Novo アッセンブリ	マッピング	マッピング	RNA-seq	miRBase ダウンロード
BLAST検索	変異検出	ピーク検出	RPKM計算	アノテーション付け
	フィルタリング	ピーク精査	群間比較	既知のmiRNAとそれ以外の分類

データ管理

データロケーション



- Genomics Workbench ではデータ保存の階層のトップをLocationと呼びます。
- デフォルトのLocationはCLC_Dataが作成されていますが、左の図のようにLocationは追加可能です。
- Location の新規追加は、Navigation Area 左上のアイコンから作成可能です。シークエンスデータはサイズが大きいため、容量が大きいディスクへLocationを作成することをお勧めします。

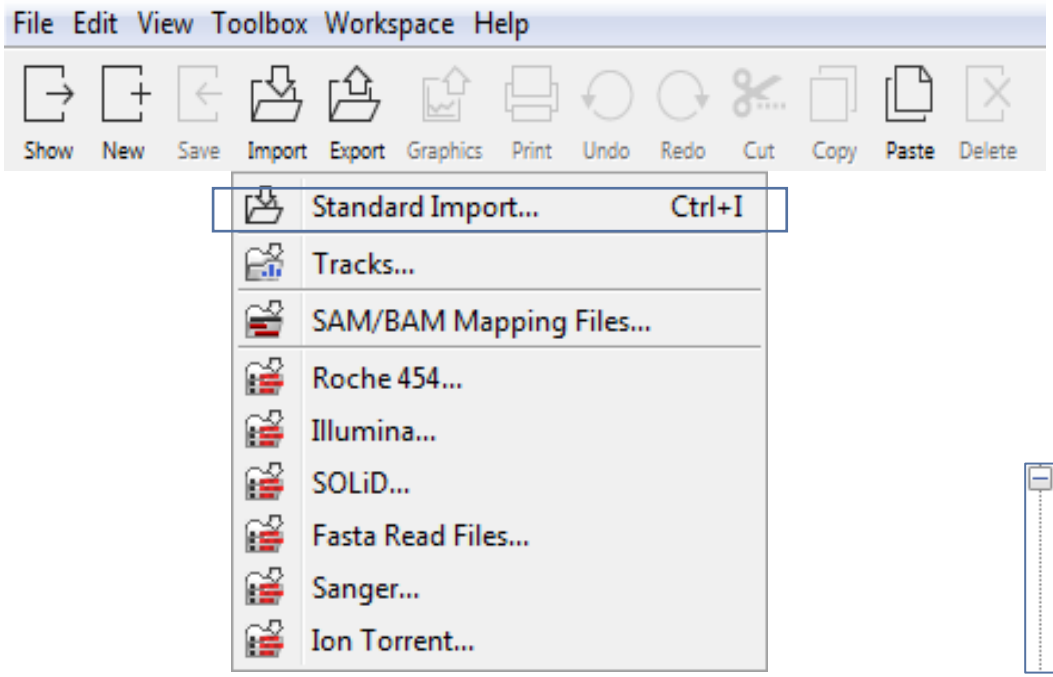


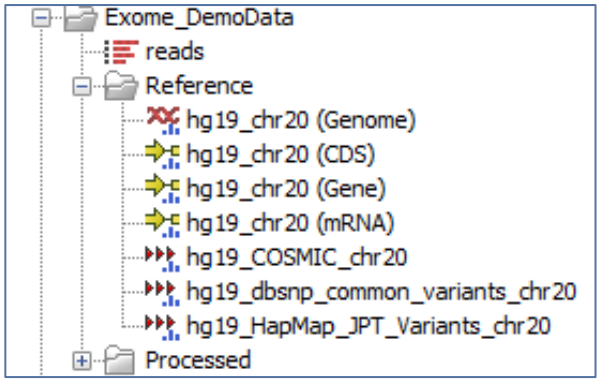
- また解析が一通り終了し、バックアップや外付けのディスクへ移動する場合は、このLocation単位での移動をお願いします。

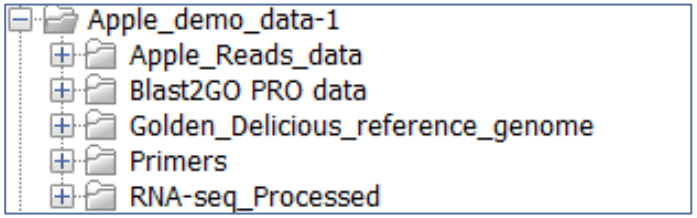
今日のデータ

データインポート

今日は、変異解析用データと、発現差解析用データを使います。それぞれzip形式で圧縮されていますが、圧縮された状態のまま、以下のImport > Standard Import よりインポートしてください。





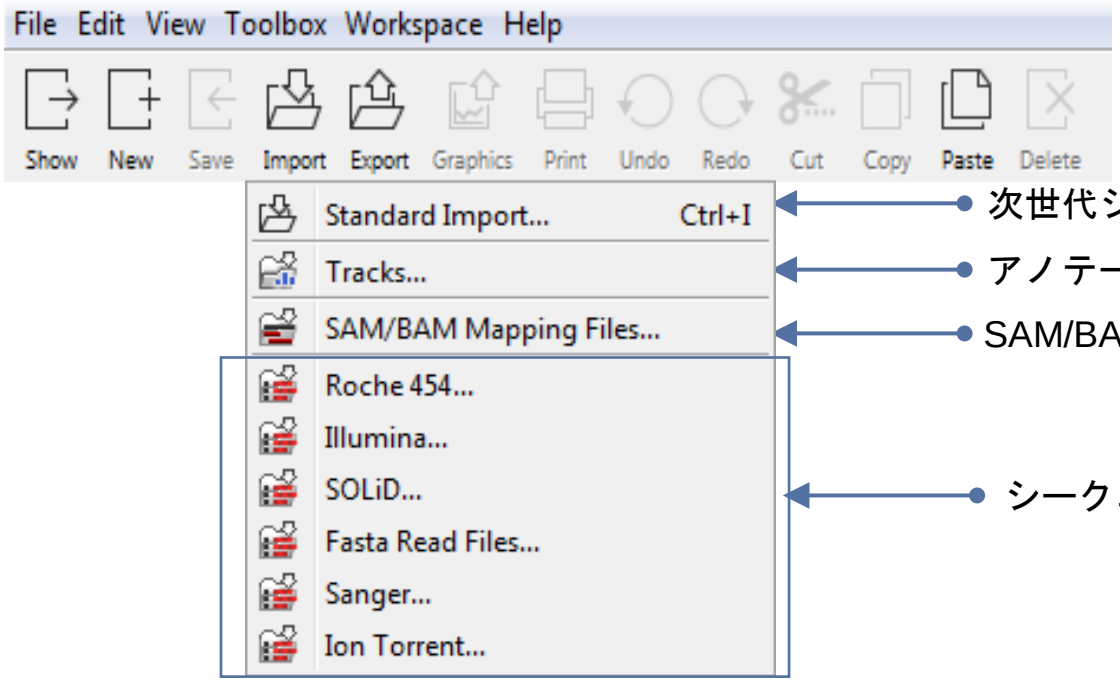


変異解析フロー

全体の流れ

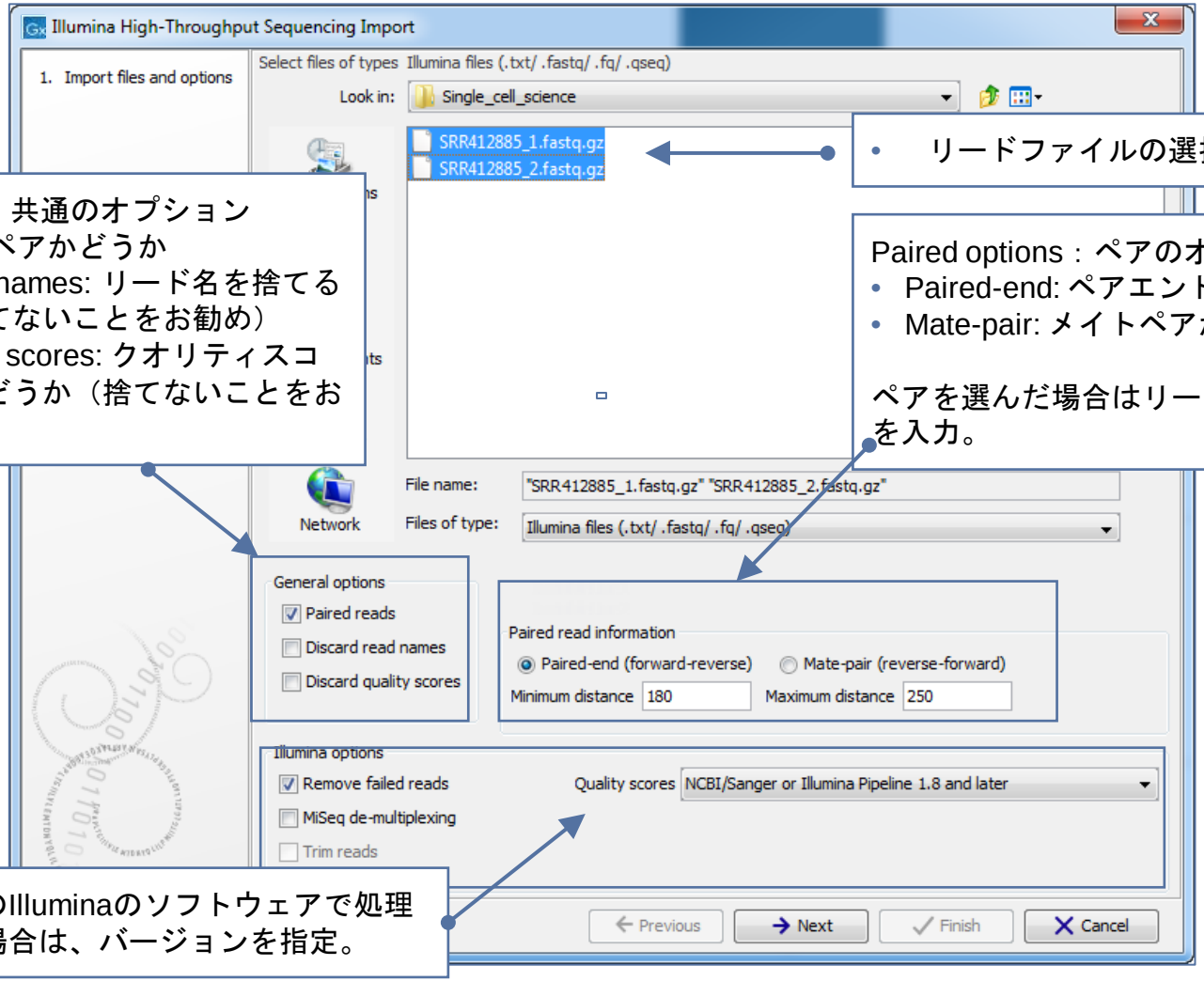
CLC Genomics Workbench
データインポート

リードデータインポート



SAM/BAMファイルは、マッピング後のデータにおいて利用される一般的なフォーマットです。

リードデータインポート：イルミナ



- リードファイルの選択

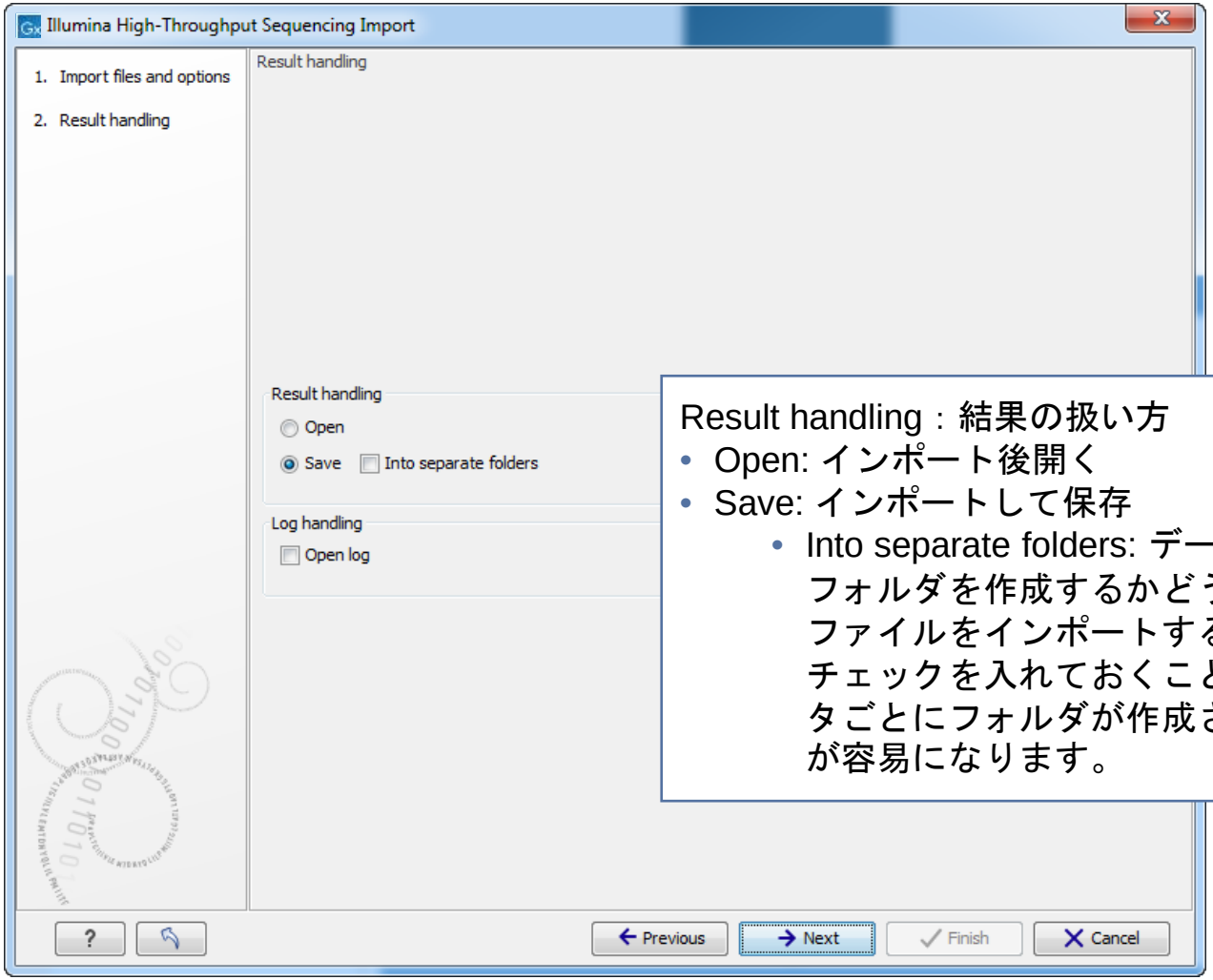
General options：共通のオプション
- Paired reads: ペアかどうか
- Discard reads names: リード名を捨てるかどうか（捨てないことをお勧め）
- Discard quality scores: クオリティスコアを捨てるかどうか（捨てないことをお勧め）

Paired options：ペアのオプション
- Paired-end: ペアエンドかどうか
- Mate-pair: メイトペアかどうか

ペアを選んだ場合はリード長を含めた距離を入力。

古いバージョンのIlluminaのソフトウェアで処理されたデータの場合は、バージョンを指定。

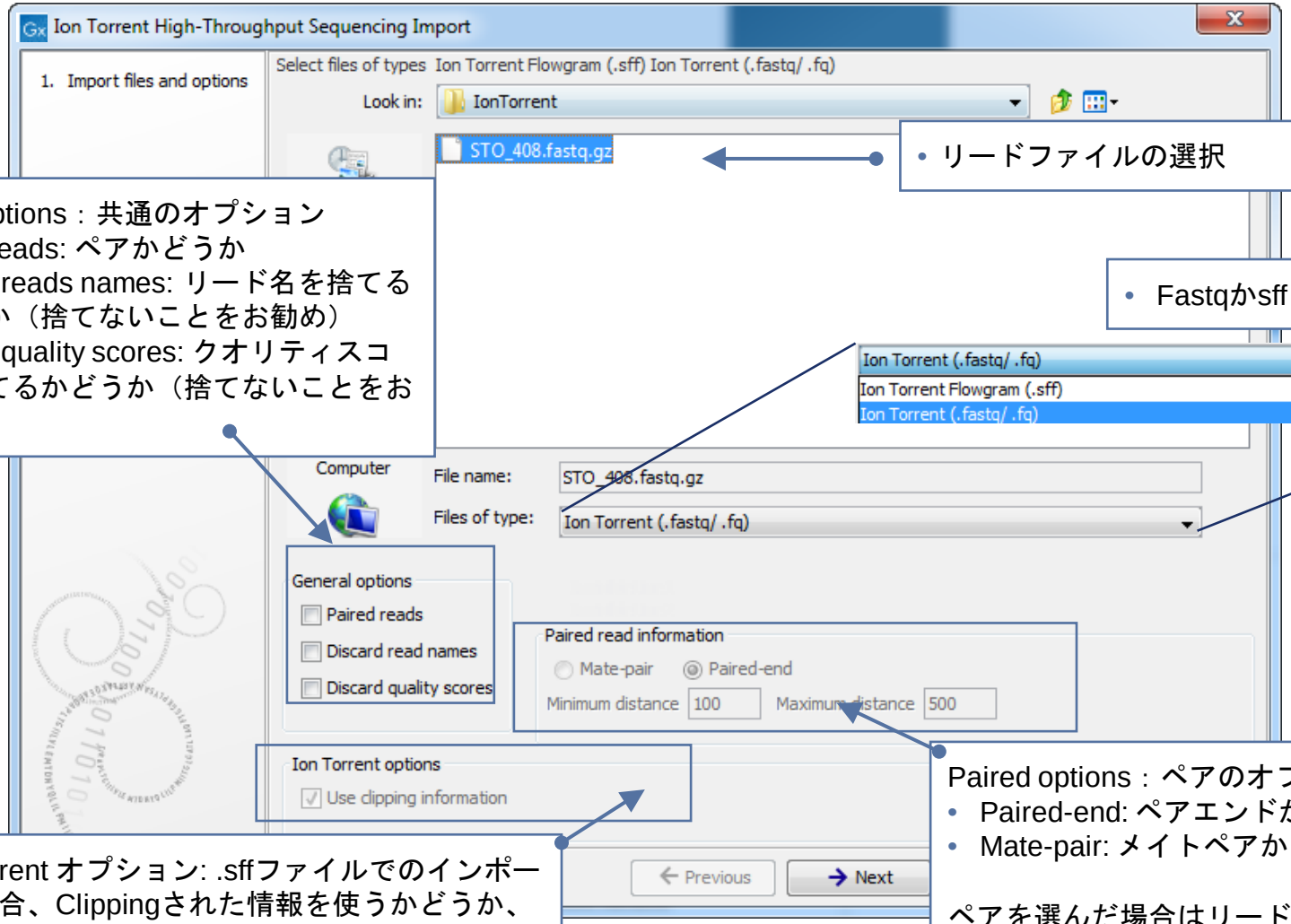
リードデータインポート：イルミナ



Result handling：結果の扱い方

- Open: インポート後開く
- Save: インポートして保存
  - Into separate folders: データごとにフォルダを作成するかどうか。複数ファイルをインポートする場合は、チェックを入れておくことで、データごとにフォルダが作成され、管理が容易になります。

リードデータインポート：Ion Torrent



- リードファイルの選択

- Fastqかsffを選択可能

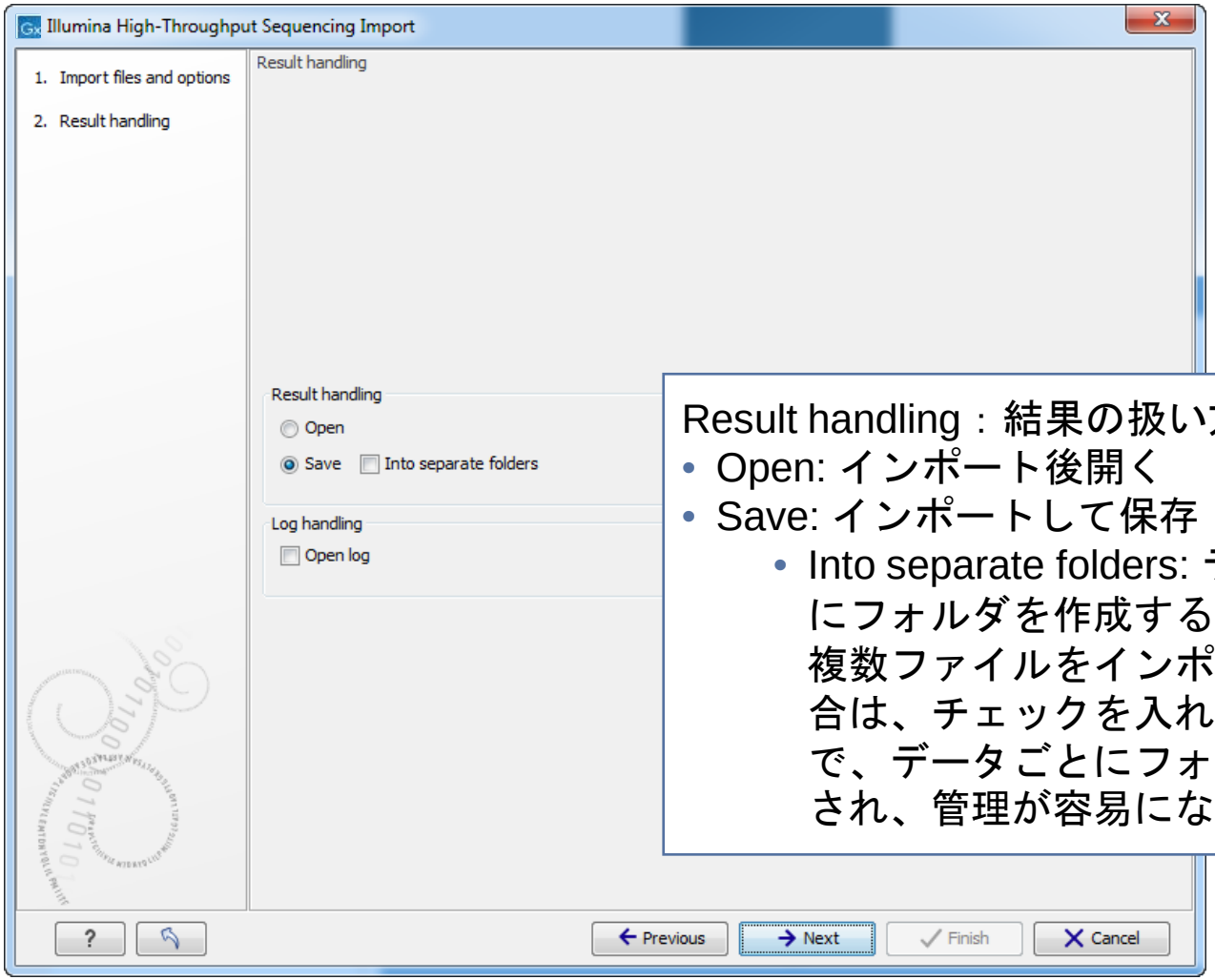
General options：共通のオプション
- Paired reads: ペアかどうか
- Discard reads names: リード名を捨てるかどうか（捨てないことをお勧め）
- Discard quality scores: クオリティスコアを捨てるかどうか（捨てないことをお勧め）

Ion Torrent オプション: .sffファイルでのインポートの場合、Clippingされた情報を使うかどうか、選択できる。

Paired options：ペアのオプション
- Paired-end: ペアエンドかどうか
- Mate-pair: メイトペアかどうか

ペアを選んだ場合はリード長を含めた距離を入力。

リードデータインポート：Ion Torrent



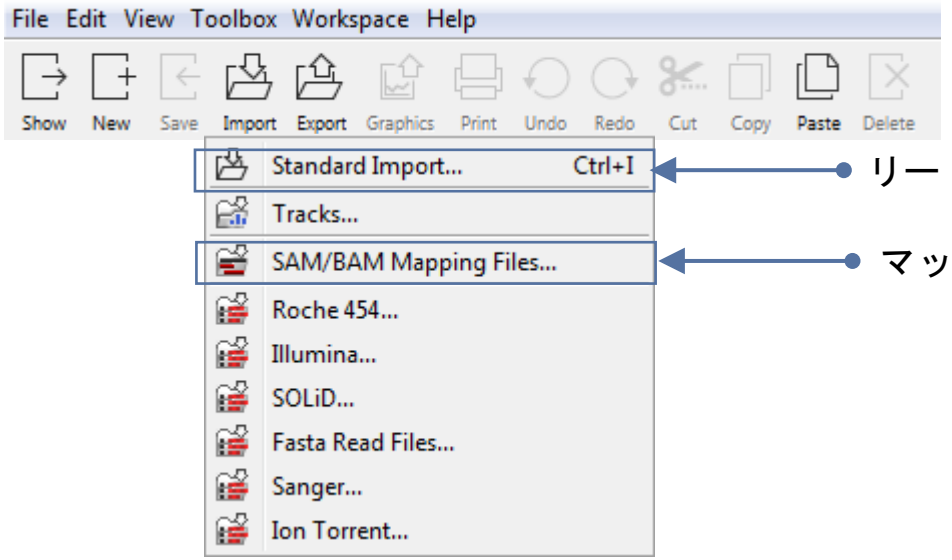
Result handling：結果の扱い方

- Open: インポート後開く
- Save: インポートして保存
  - Into separate folders: データごとにフォルダを作成するかどうか。複数ファイルをインポートする場合は、チェックを入れておくことで、データごとにフォルダが作成され、管理が容易になります。

データインポート

リードデータインポート：Ion Torrent (Unmapped BAMファイル)　※注意

Ion Torrentのシークエンサーデータを処理するTorrent Suitでは、バージョン3.0以降、デフォルトでは、fastqファイルやsffファイルが作成されず、Unmapped BAM ファイルが作成されます。Unmapped BAM ファイルは、Import > Standard Import よりインポートいただくことで、fastqファイルをインポートした場合と同じようにインポートが可能です。

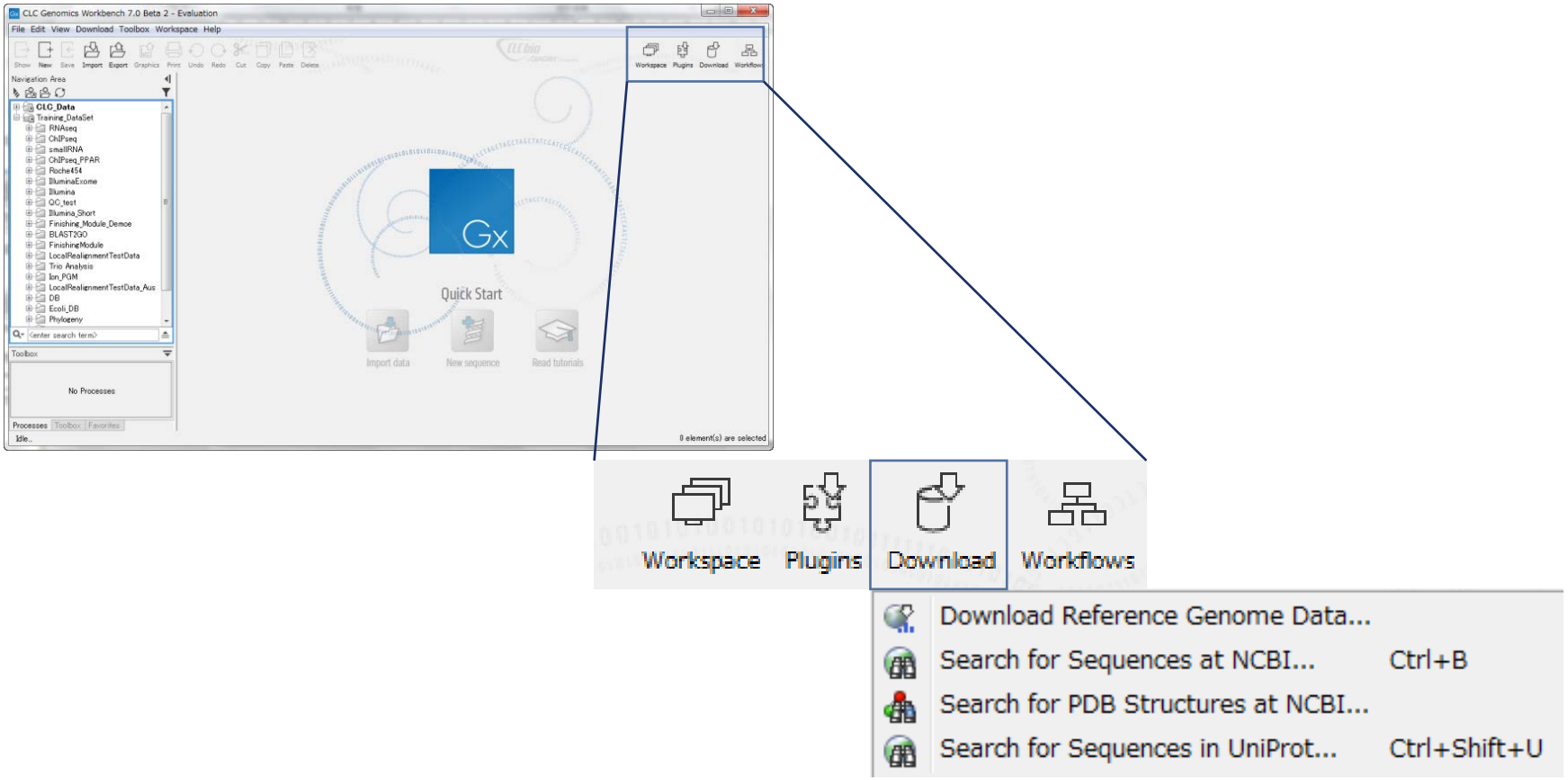


リードデータとしてインポートされます。

マッピングデータとしてインポートされます。

ゲノムインポート

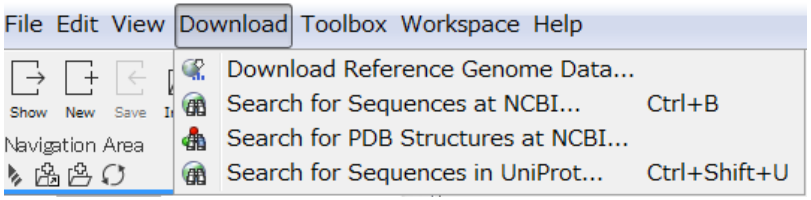
ゲノムデータは、よく知られているモデル動物についてはのDownload Genome よりインポートできます。



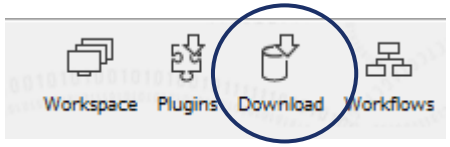
データインポート

ゲノム配列の入手方法

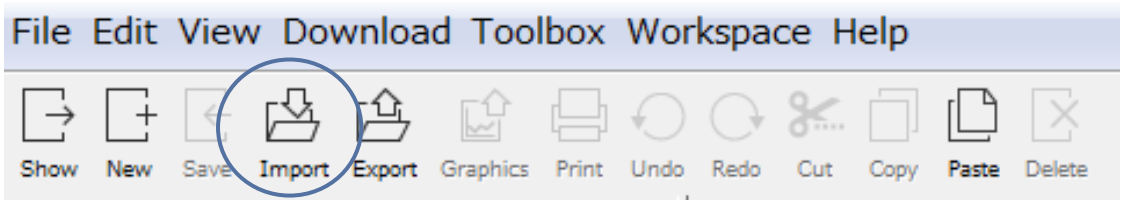
- Download 機能を用いる



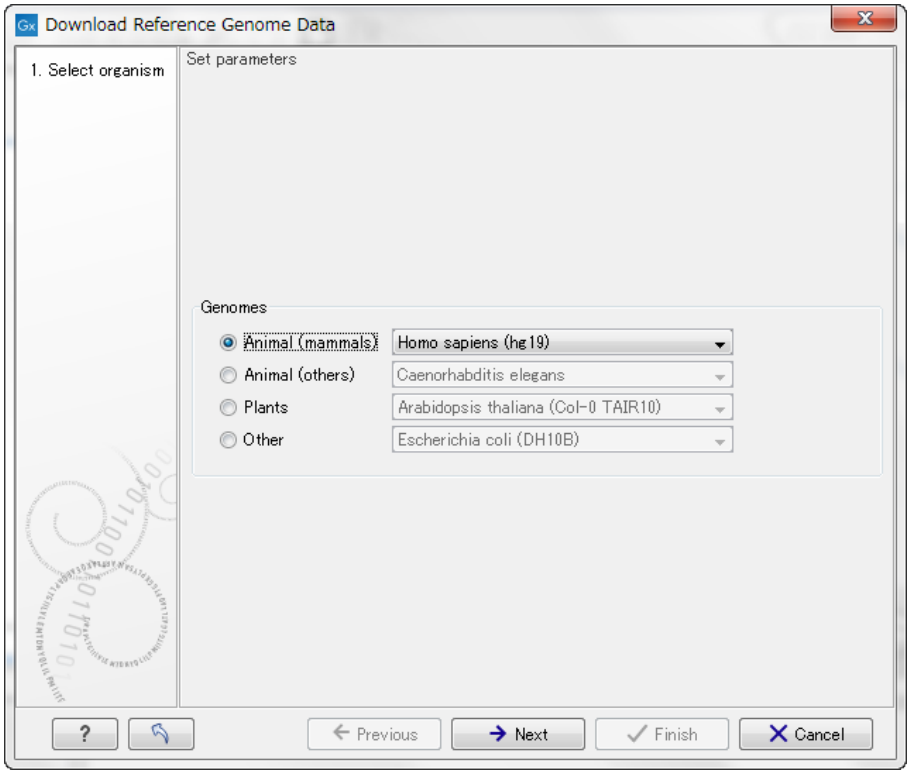
または



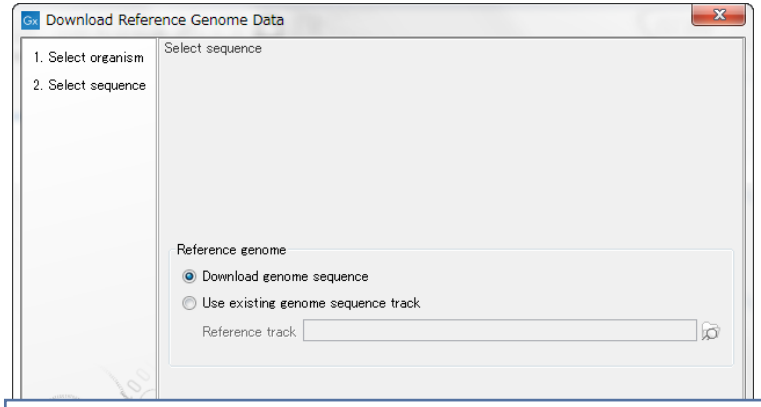
- Download サイトからダウンロードしたファイルをインポートする



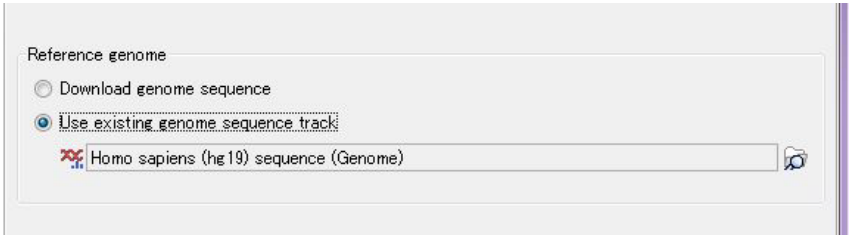
ゲノムインポート



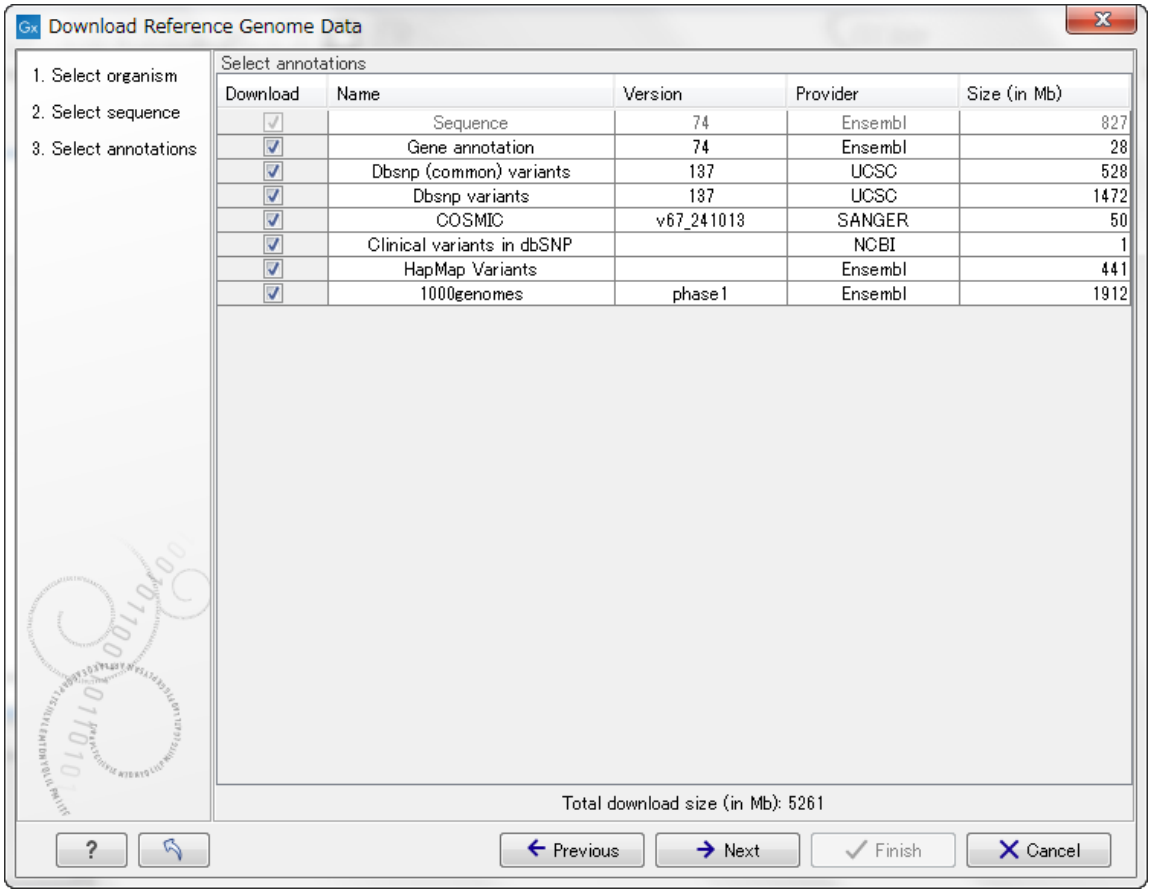
- ドロップダウンリストから生物種を選択。



- Download genome sequence: 新規にゲノムをダウンロードする場合。
- Use exsting genome sequence track: すでにダウンロードしたゲノムにアノテーションを追加する場合。以下のようにトラックのフォーマットになっているゲノムを選択。



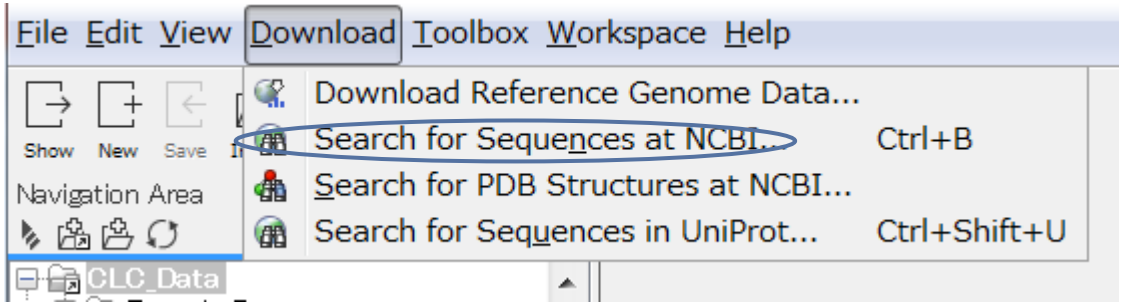
ゲノムインポート



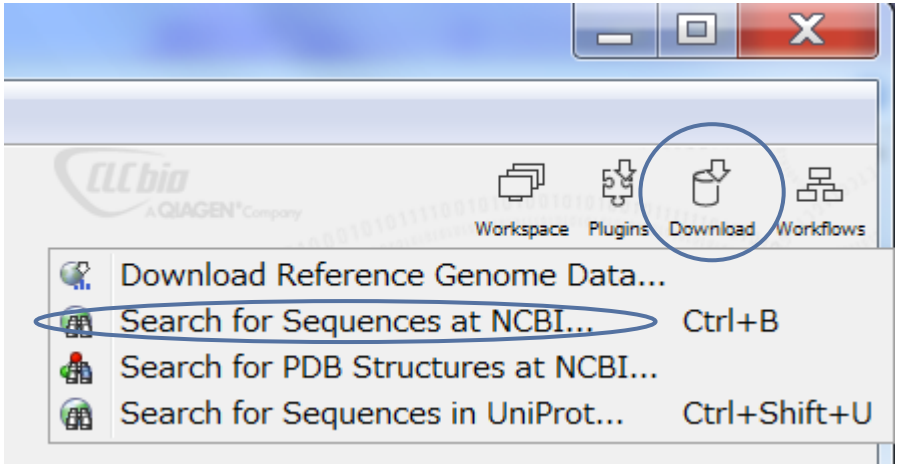
- 希望するアノテーションにチェックを入れる。ゲノム配列をダウンロードするときは、Sequences にもチェックを入れる。
- 選択した生物種により、表示されるアノテーションの種類は異なります。

データインポート

NCBIで検索してインポート



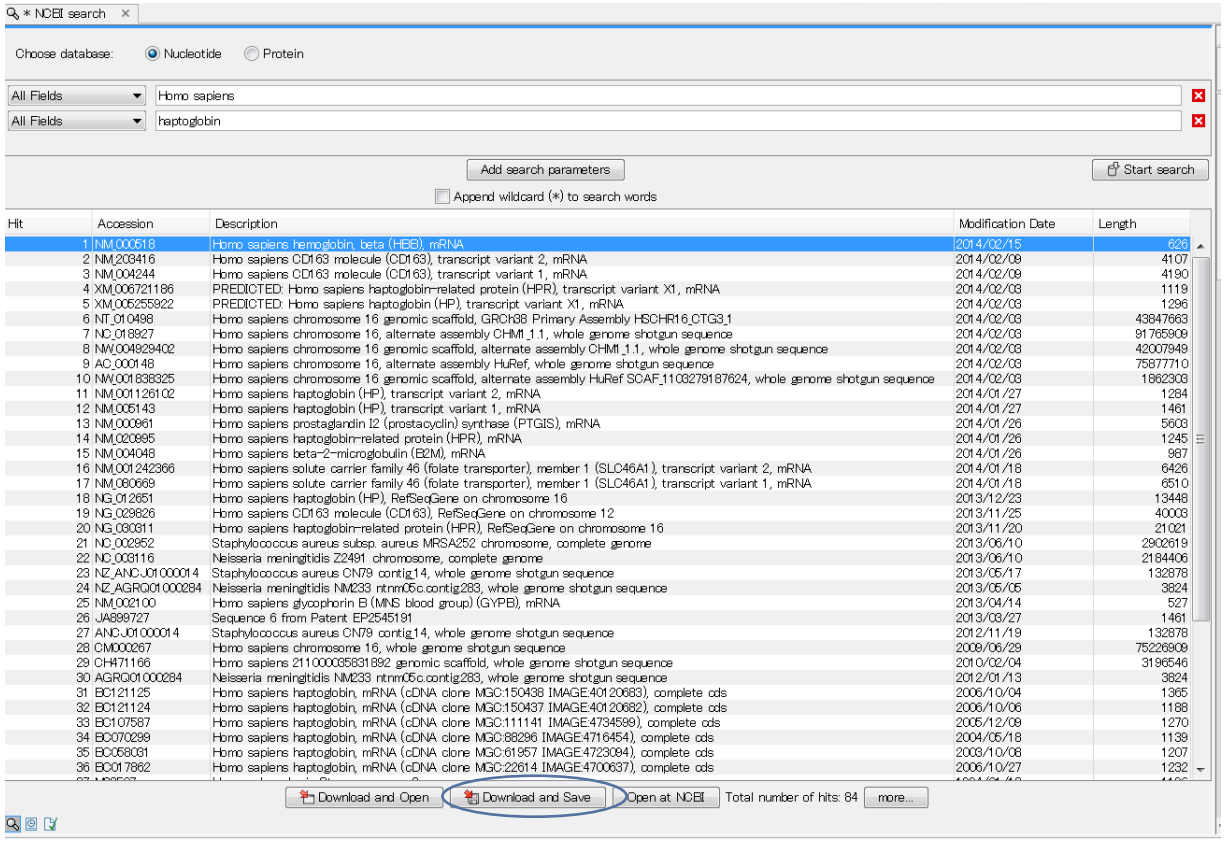
または



- NCBI のサイトに検索をかけて、直接ゲノム配列をダウンロードすることができます

データインポート

Search for Sequences at NCBI



- 検索のキーワードを入れて、Start search をクリックします
- 目的の配列を選択して、Download and Save で配列をダウンロードできます

データインポート

アノテーションインポート

- Download Genome 以外にも、アノテーションファイルをインポート可能です。
- アノテーションとして取り込めるファイルは以下のフォーマットです。
- アノテーションファイルをインポートする際には、対象となるゲノム配列がすでにインポートされ、Trackのフォーマットになっていることが前提です。

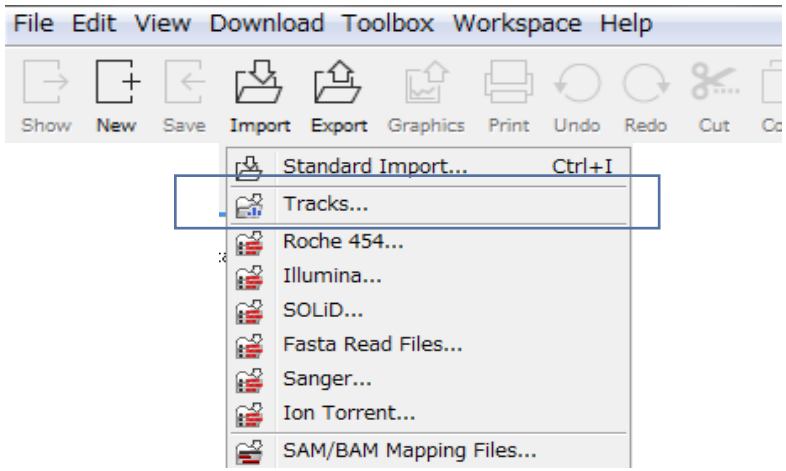
  - VCF
  - GFF/GTF/GVF
  - BED
  - Wiggle
  - Complete Genomics Var file
  - UCSC Variation table damp
  - COSMIC variation database

※変異のデータについても、アノテーションとして自分の変異へアノテーションとして情報の追加や比較ができるため、アノテーションのインポート可能フォーマットに含めています。

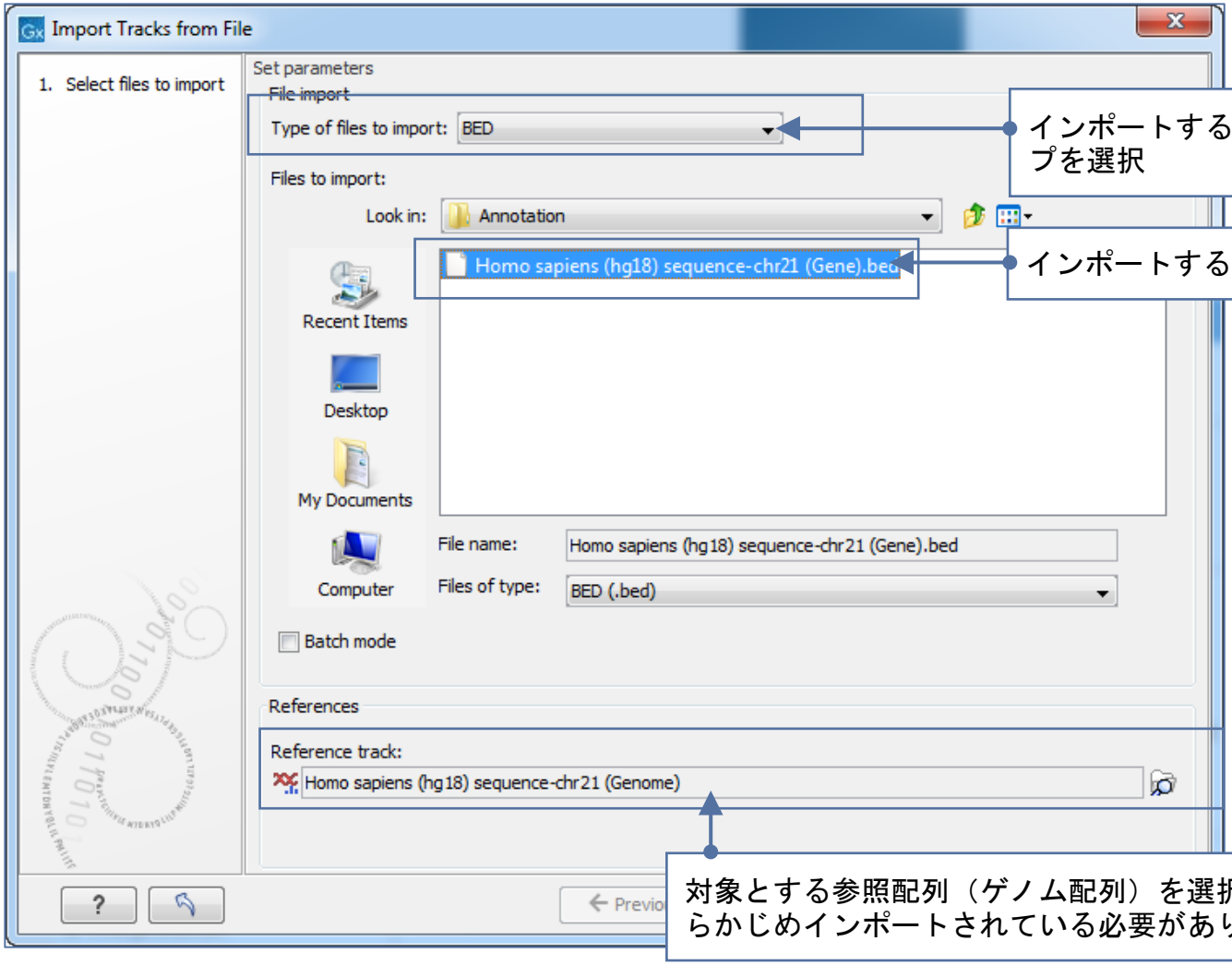
データインポート

アノテーションインポート

アノテーションのインポートは、Import > Tracks より行います。



トラックインポート

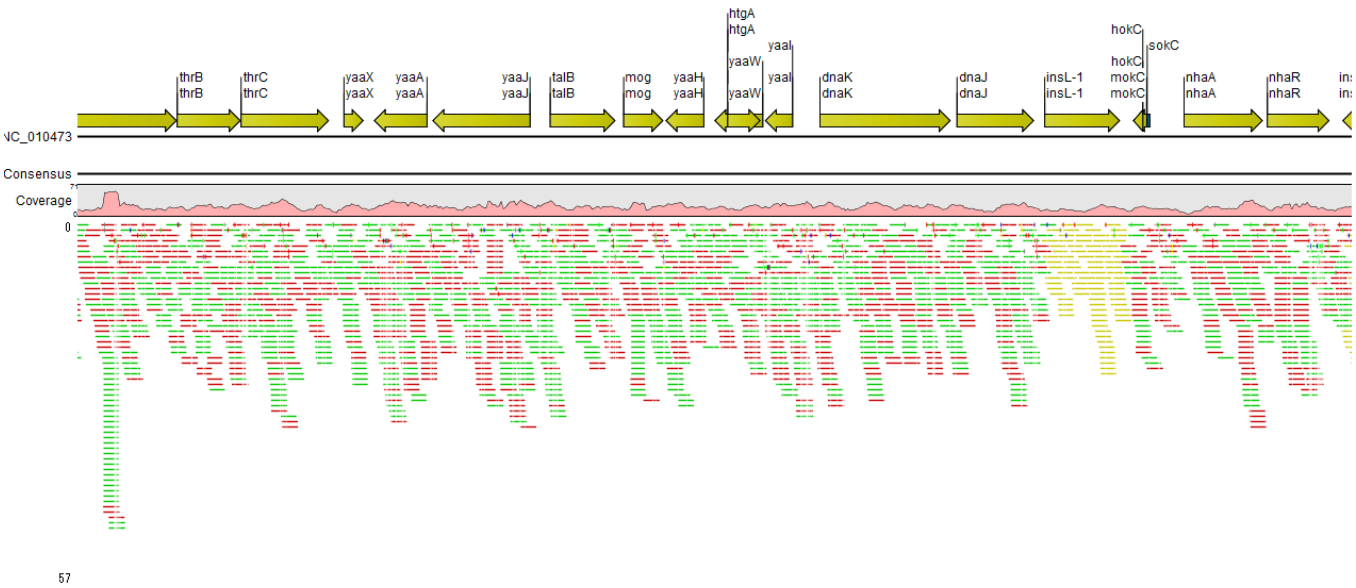

Import Tracks from File
1. Select files to import
Set parameters
File import
Type of files to import: BED
Files to import:
Look in: Annotation
Homo sapiens (hg18) sequence-chr21 (Gene).bed
Recent Items
Desktop
My Documents
Computer
File name: Homo sapiens (hg18) sequence-chr21 (Gene).bed
Files of type: BED (.bed)
Batch mode
References
Reference track:
Homo sapiens (hg18) sequence-chr21 (Genome)
インポートするファイルのタイプを選択
インポートするファイルを選択
対象とする参照配列（ゲノム配列）を選択。あらかじめインポートされている必要があります。

データフォーマット編
スタンドアロンフォーマットと トラックフォーマット

スタンドアロンフォーマット

スタンドアロンフォーマットでは、1つのデータに配列情報、アノテーションがセットになっています。

reads mapping

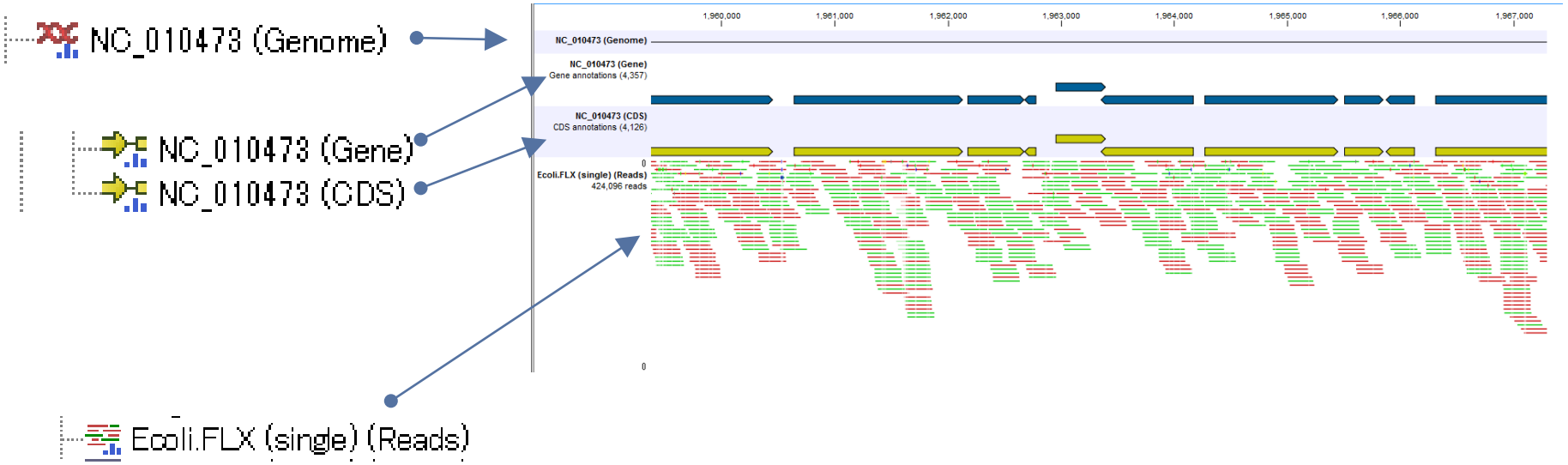


トラックフォーマット

トラックフォーマットでは、リードやゲノム配列、アノテーションがばらばらのファイルになっており、好きに組み合わせて表示が可能です。

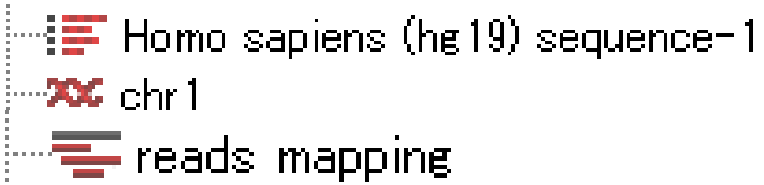
トラックリスト

複数のトラックを組み合わせてTrack list を作ることで好きなビューを作成できます。



フォーマットとアイコン表示

スタンドアロンフォーマット



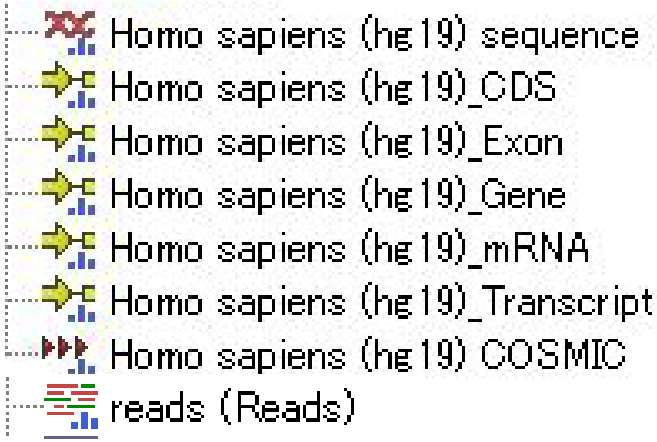
染色体のセットやリード配列など配列のセット

染色体1本など1つの配列

リードマッピング

トラックフォーマット

青いヒストグラムが目印



ゲノムTrack

アノテーションTrack

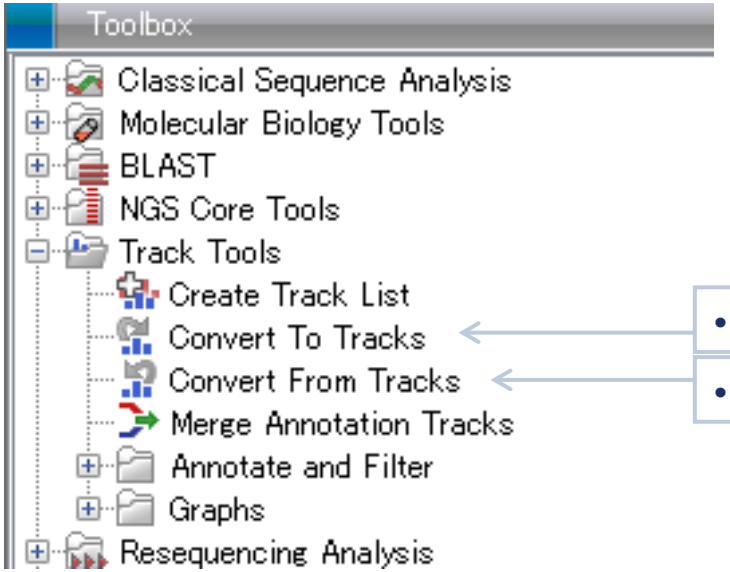
変異Track

リード（マッピング）Track

解析によって必要とするフォーマットが異なります。
スタンドアロン⇔トラックの変換は自由に行えます。

フォーマットの変換

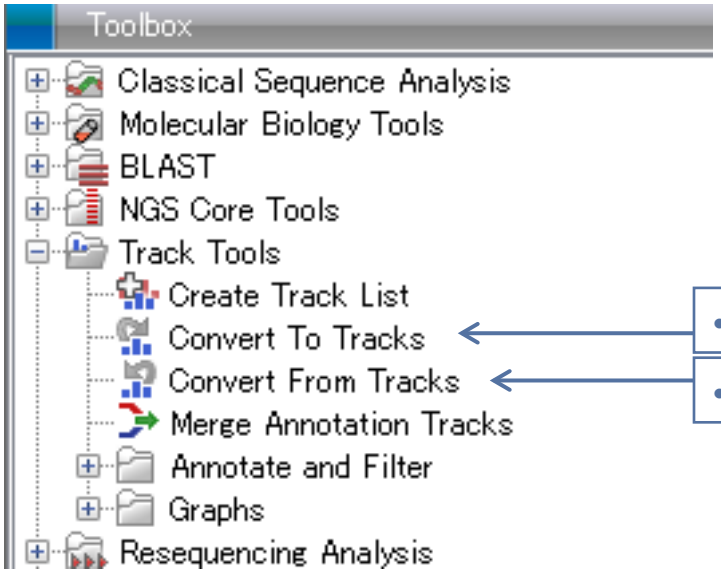
トラックフォーマットからスタンドアロンフォーマット、またスタンドアロンフォーマットからトラックフォーマットへは、 Toolbox > Track tools の中のツールを使って変換可能です。



- スタンドアロンフォーマットからトラックへの変換。
- トラックからスタンドアロンフォーマットへの変換。

フォーマットの変換

トラックフォーマットからスタンドアロンフォーマット、またスタンドアロンフォーマットからトラックフォーマットへは、 Toolbox > Track tools の中のツールを使って変換可能です。

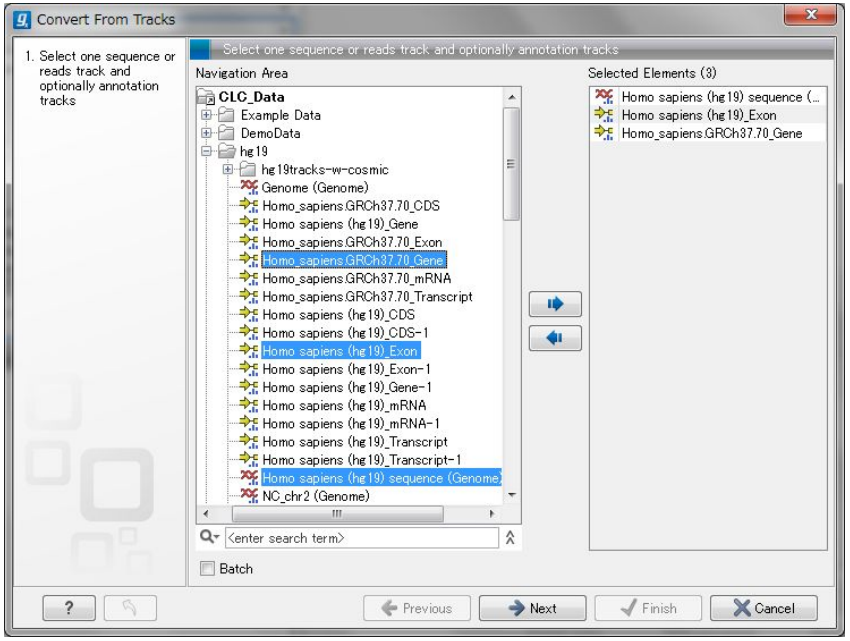


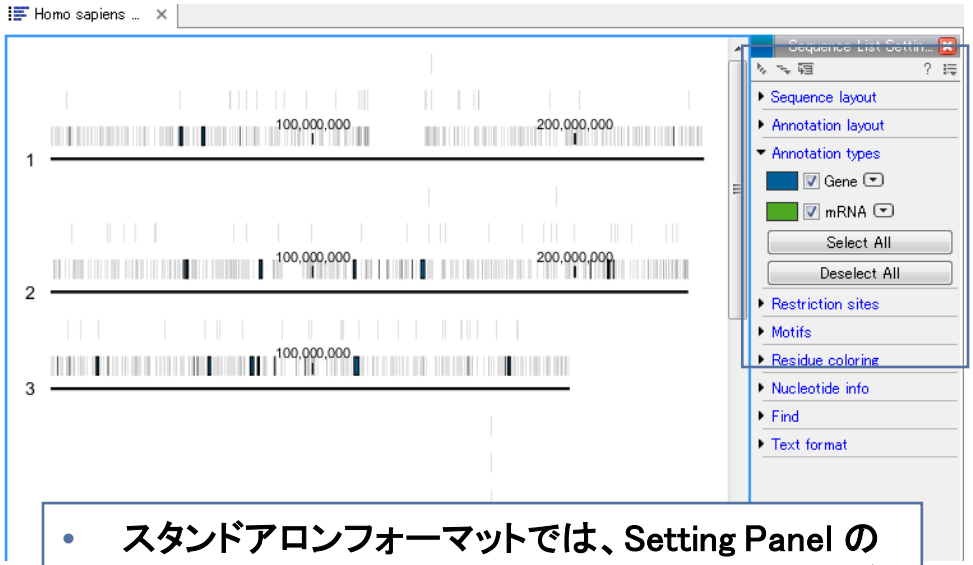
- スタンドアロンフォーマットからトラックへの変換。
- トラックからスタンドアロンフォーマットへの変換。

スタンドアロンフォーマットへ変換する場合、スタンドアロン内に含めるアノテーショントラックを含めて変換するようにしてください。

フォーマットの変換

スタンドアロンフォーマットへ変換する場合、スタンドアロン内に含めるアノテーショントラックを含めて変換するようにしてください。





- スタンドアロンフォーマットでは、Setting Panel の Annotation Type からどういったアノテーションが付属しているか確認できます。

クオリティチェックとトリミング

クオリティチェックとトリミング

Quality Report作成: Create Sequencing QC Report

- インポートしたリードのクオリティがどのぐらいか、その後のトリミングや、PCR Duplicate の状況などを確認するためにレポートを作成。

トリミング: Trim Sequences

- アダプターの除去、クオリティスコアによる除去、長さを指定した除去などを選択・組み合わせてトリミング。

上記処理の後に再度Quality Reportを作成すると処理前と処理後でのリードのクオリティを比較でき、便利です。

クオリティスコア

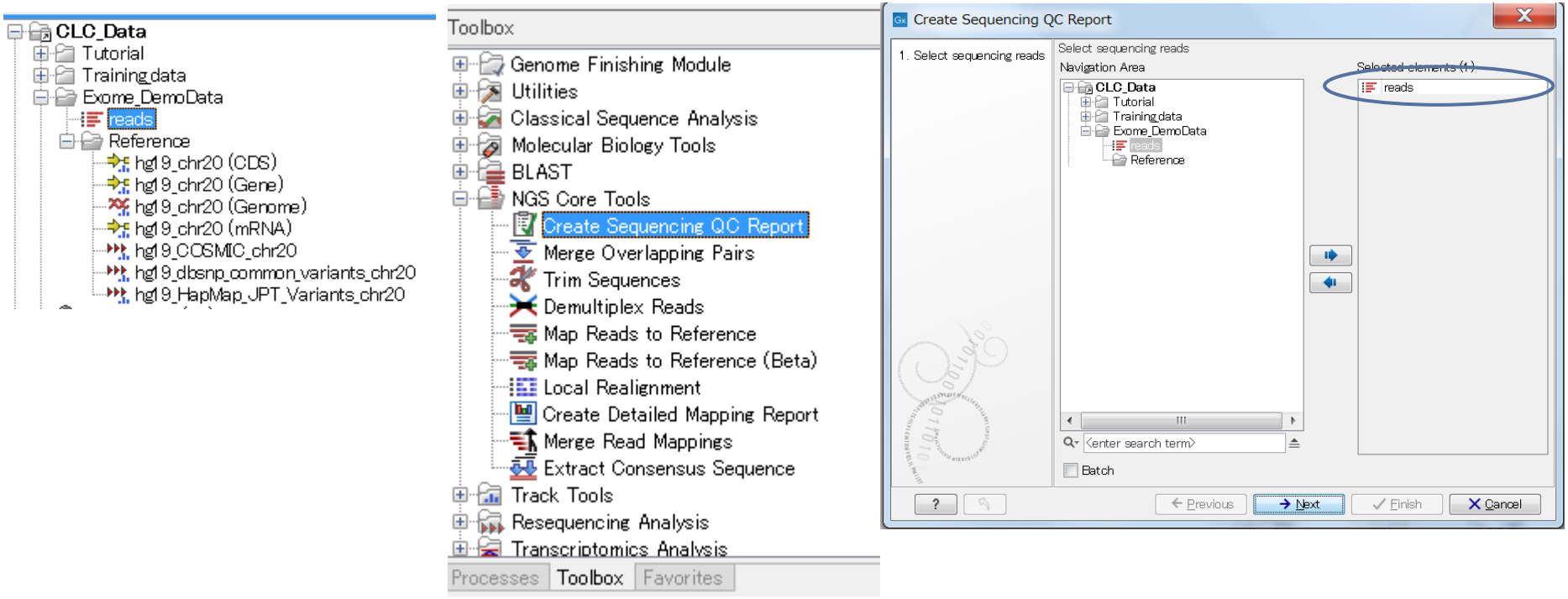
シークエンサーから出てきたリードは、各塩基ごとにエラーの確率の値を持っている。

Genomics Workbench へインポートされた時点で、Phred Score に変換されるようになっています。Pred Score は、塩基のエラー確率のLogを取り、-10をかけてスコア化したものです。値が大きくなるほど精度が高いことをあらわしています。

$$PhredScore = -10\log_{10} P_{err}$$

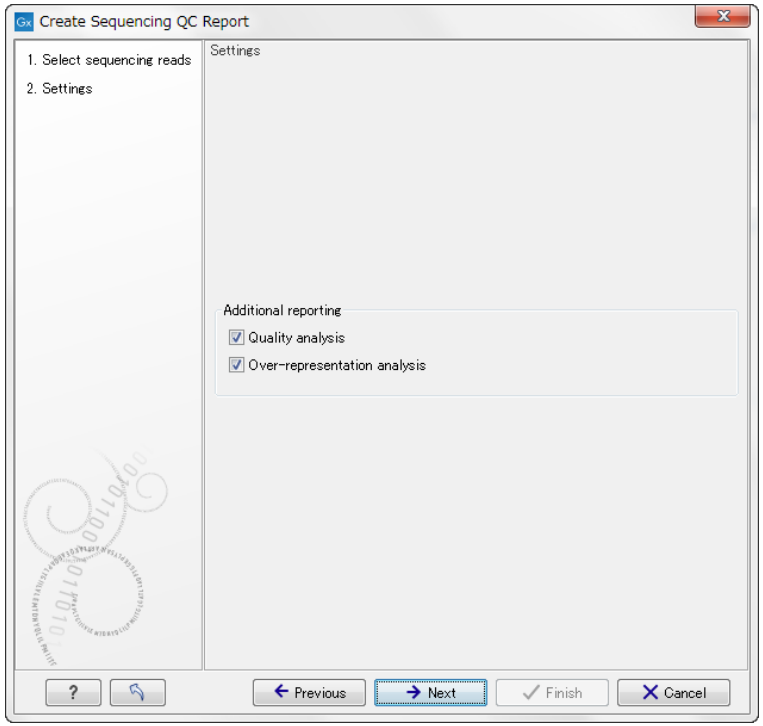
Phred Score	Error の確率	Base call の精度
10	1/10	90%
20	1/100	99%
30	1/1,000	99.9%
40	1/10,000	99.99%
50	1/100,000	99.999%
60	1/1,000,000	99.9999%

QCレポート作成：Create Sequencing QC Report

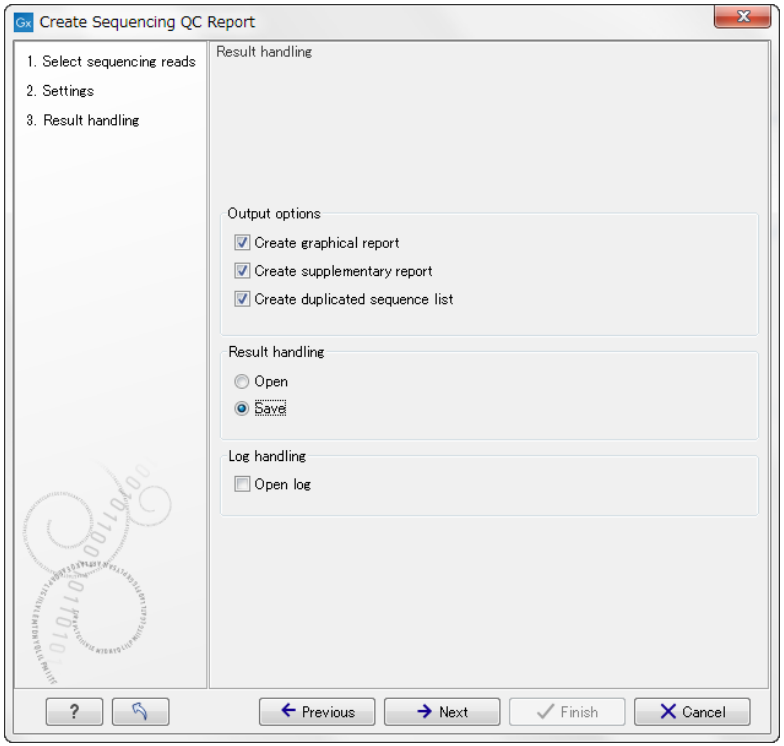


- Navigation Areaから使用するリードデータを選択。
- Toolboxから NGS Core Tools > Create Sequencing QC Report を選択、ダブルクリック。
- ウィザードが起動し、選択したデータが選ばれていることを確認。

QCレポート作成：Create Sequencing QC Report

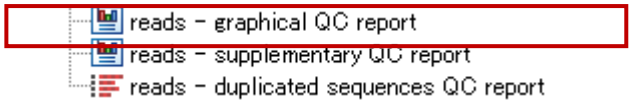


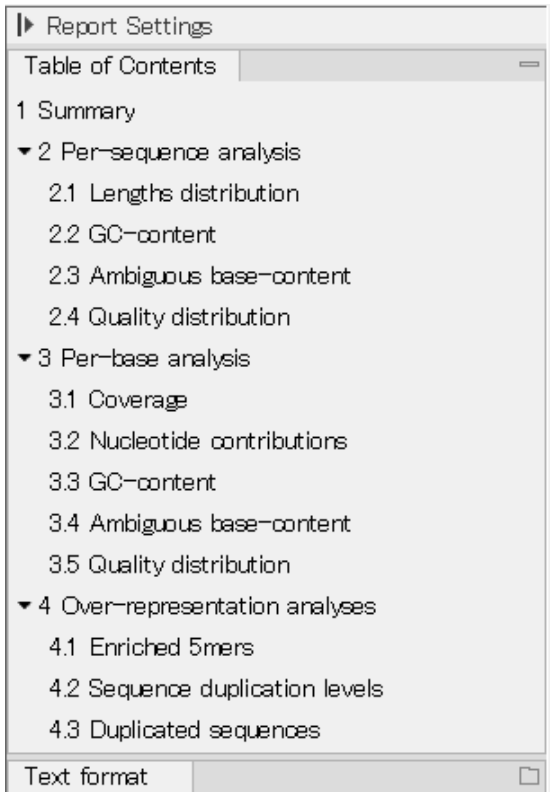
- Quality analysis: クオリティスコアに関する解析。
- Over-representations analysis: 過度に現れているような塩基配列などの解析。



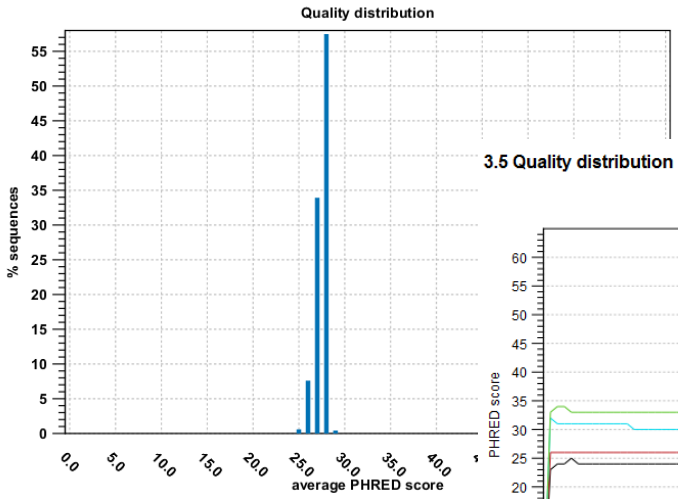
- Create graphical report: グラフィカルなレポート作成。
- Create supplementary report: 数値のレポート作成。
- Create duplicated sequence list: 重複のあった配列のリスト作成。

QCレポート作成：Create Sequencing QC Report

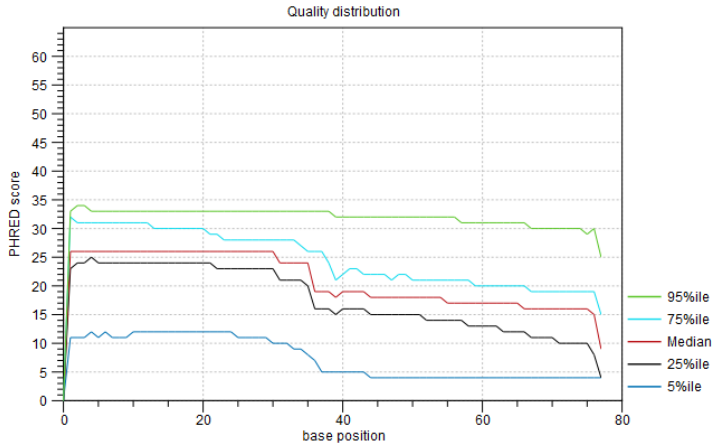




2.4 Quality distribution



3.5 Quality distribution



Base-quality distribution along the base positions.

トリミング原理

3種類のトリミング

アダプター除去	• あらかじめ登録されているアダプターの除去 • 新規で独自の配列を登録することも可能
クオリティトリミング	• Quality Score を使い、Quality の低い配列が連続するようになる箇所からカット • 正確に読めていない塩基をいくつ許容するか
長さによる除去	• 塩基数を指定して、5末端、3末端をカット • Quality Scoreでカット後、短くなりすぎた配列をカット

クオリティスコア

Trimming ではQuality Score を使い、累積のQuality Score がある一定の値より大きいものが続いた場合に、その箇所を取り除く、という処理を行います。

具体的には以下：

1. Phred Score をp値へ変換

$$P_{err} = 10^{-\frac{PhredScore}{10}}$$

2. Trimming 中に設定するパラメータ（Limit）とp値の差を計算
3. 差の累積和を計算。このとき、0以下の値は0とする
4. Trimming後のリード開始点は累積和がはじめて0以上になった点。Trimming後のリード終了点は累積和が最大の点

原理

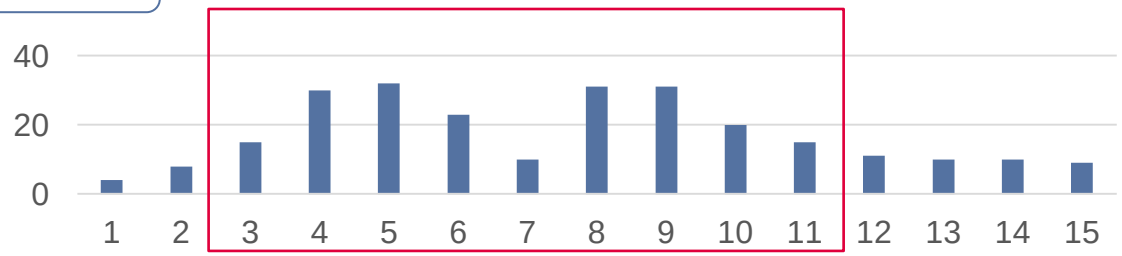
リード配列	G	C	C	C	A	T	G	T	T	C	G	A	T	G	C
Phred score	4	8	15	30	32	23	10	31	31	20	15	11	10	10	9
p値	0.40	0.16	0.03	0.00	0.00	0.01	0.10	0.00	0.00	0.01	0.03	0.08	0.10	0.10	0.13
Limit - p値 (D)	-0.35	-0.11	0.02	0.05	0.05	0.04	-0.05	0.05	0.05	0.04	0.02	-0.03	-0.05	-0.05	-0.08
(D)の累積和	0.00	0.00	0.02	0.07	0.12	0.16	0.11	0.16	0.21	0.25	0.27	0.24	0.19	0.14	0.06

Limit = 0.05の場合

スタート点：
累積和が0より大きくなった塩基

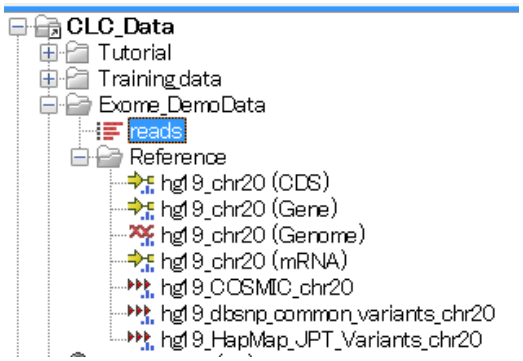
終了点：
累積和が最大を示す塩基

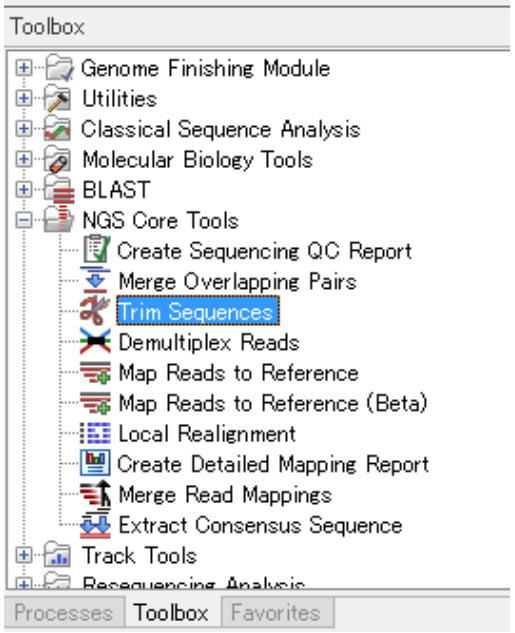
Phred score の棒グラフ

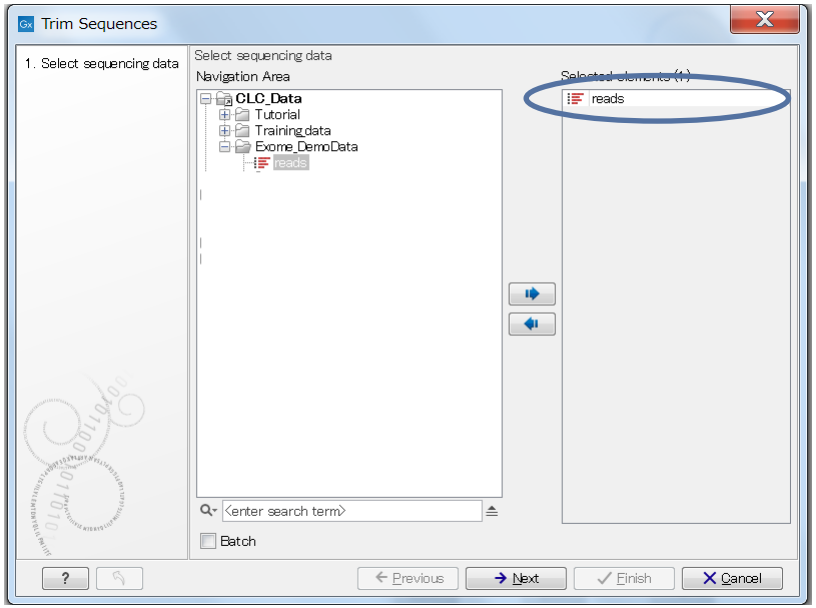


グラフより、ある程度クオリティが高くなった場所からリードを使い、クオリティが連続して悪くなっている箇所からリードをトリムしていることがわかる。
※途中、1塩基のみクオリティが低いような場合は、必ずしもトリムされない。これはできるだけリードを長く保とうとするため。

トリミング



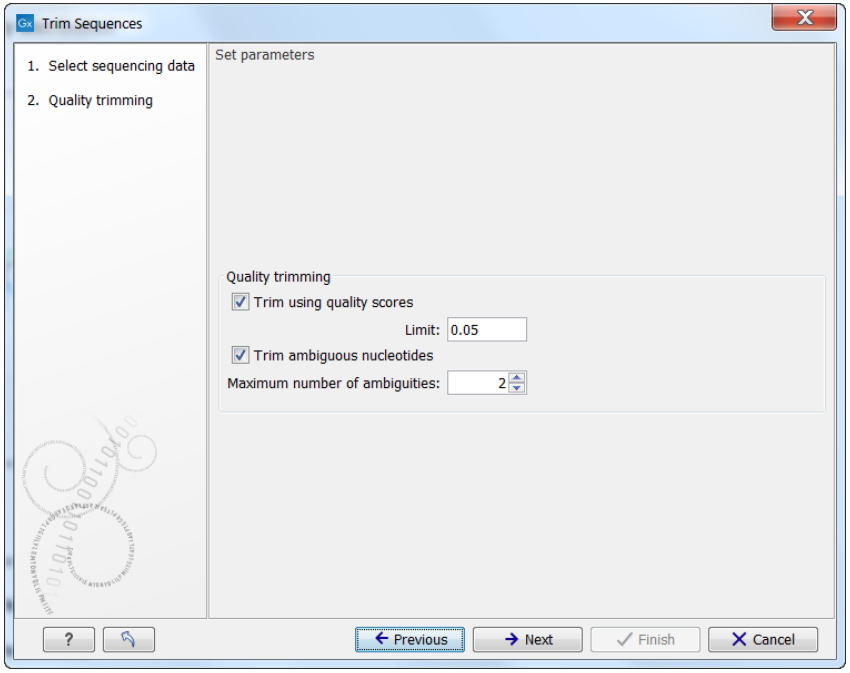


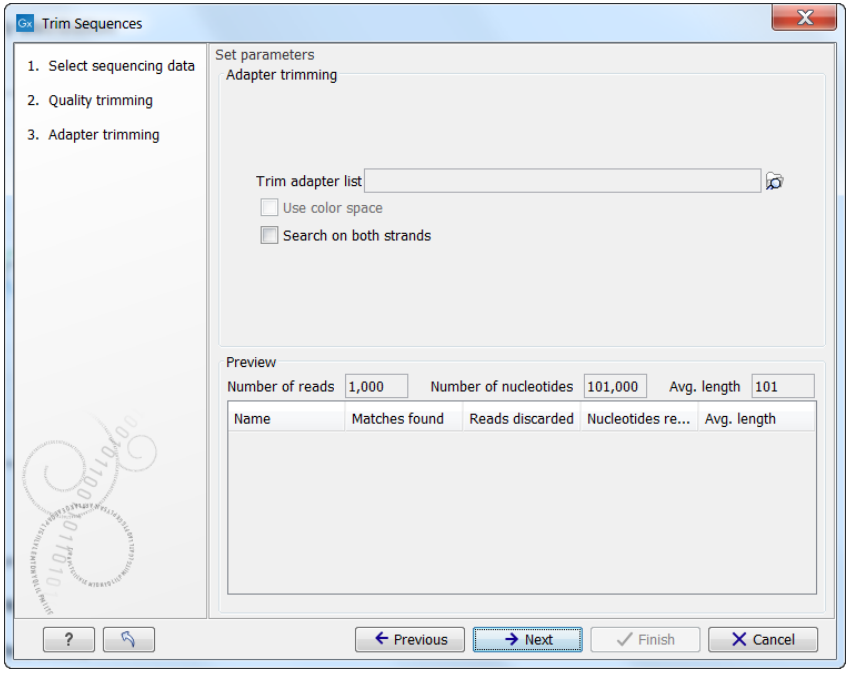


- Navigation Areaから使用するデータを選択。
- Toolboxから Trim Sequences を選択、ダブルクリック。
- ウィザードが起動し、選択したデータが選ばれていることを確認。

トリミング

クオリティトリミング



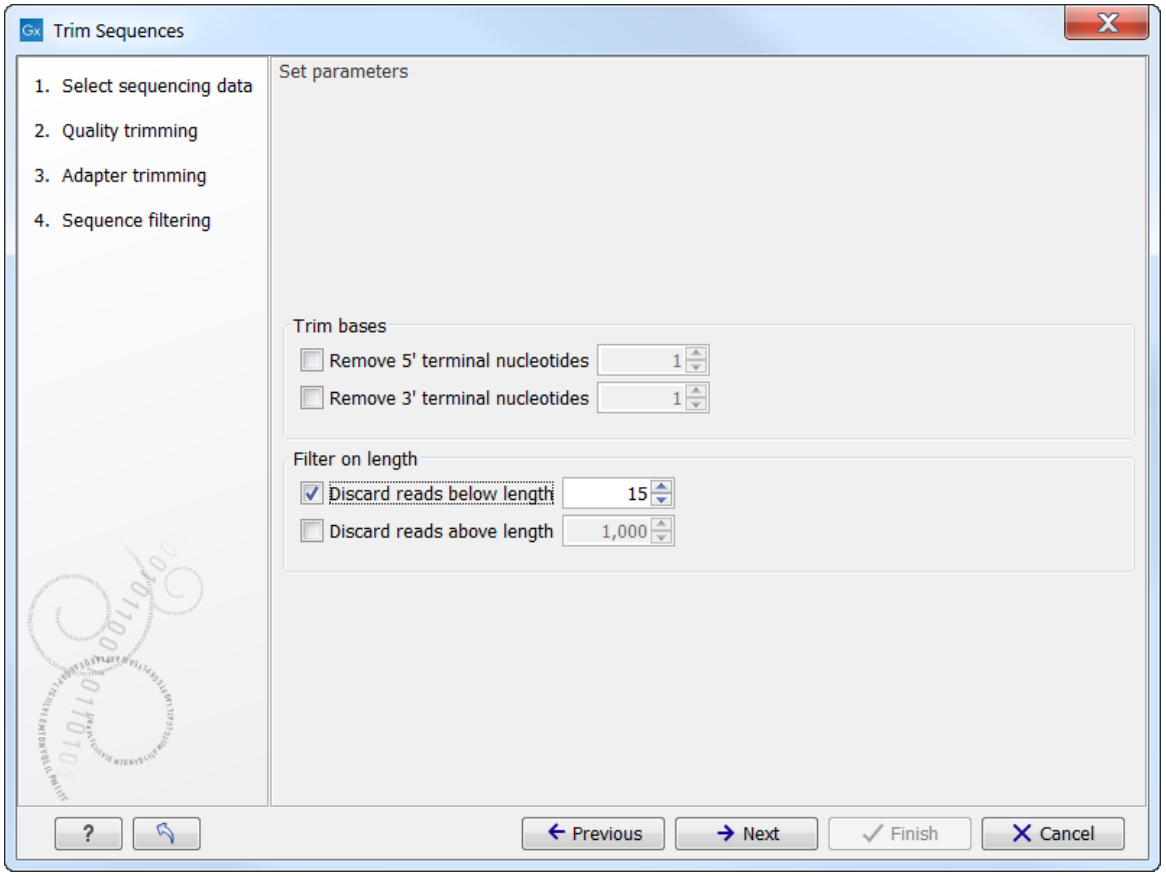


- Trim using quality scores ： トリミングに使用するLimitパラメータを決定
- Trim ambiguous nucleotides：N表示される塩基について、最大何塩基まで保持させるか。

- 今回はアダプターは設定なし。

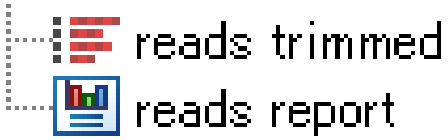
長さによるトリミング

- 5末、3末の塩基数を指定してカットする
- Quality Scoreによるトリミングであまりに短いリードの除去など長さによるトリミング

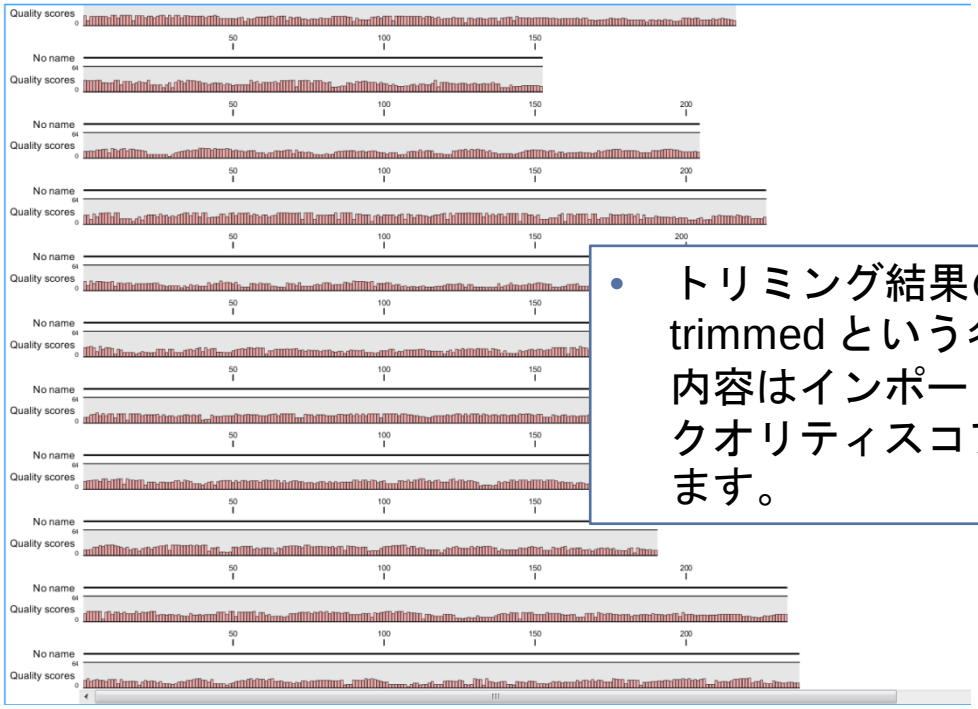


トリミング結果

結果



- トリミング後は、トリムされたリードと、レポートを作成した場合は、そのレポートが作成されます。



- トリミング結果のデータはファイル名の後に trimmed という名前が付いています。ファイル内容はインポート後のデータ同様に、配列と、クオリティスコアを含んだファイルとなっています。

結果

1 Trim summary

Name	Number of reads	Avg.length	Number of reads after trim	Percentage trimmed	Avg.length after trim
s_1_1_sequence (paired)	5,244,764	35.0	5,244,081	99.99%	34.5

2 Read length before / after trimming

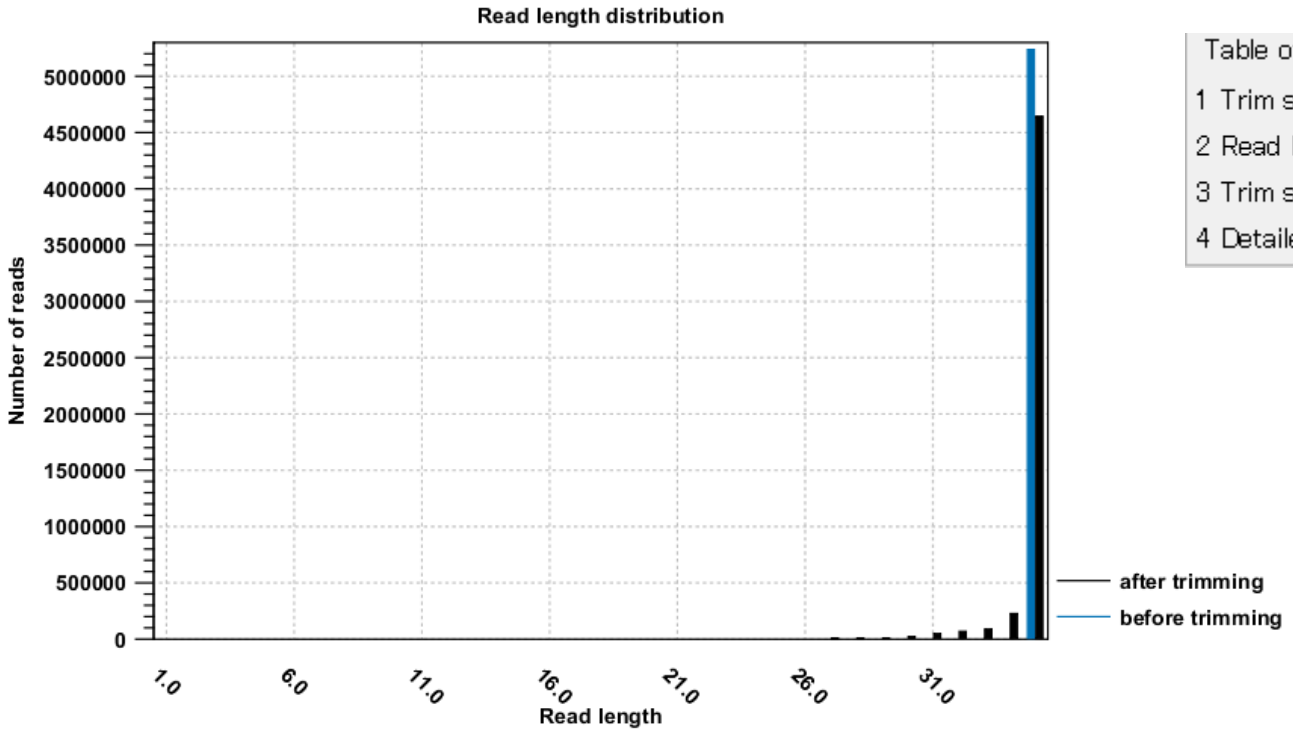


Table of Contents

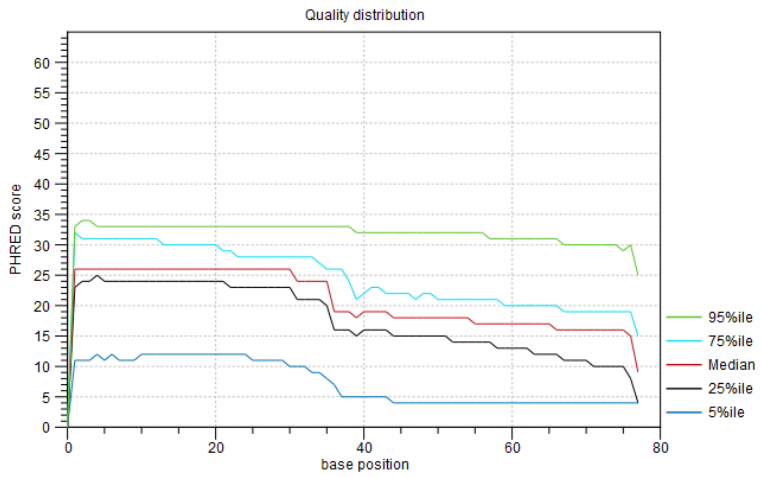
QCレポート　再作成による比較

エクササイズ

- トリミング後のデータでレポートを作成してみましょう！

Before

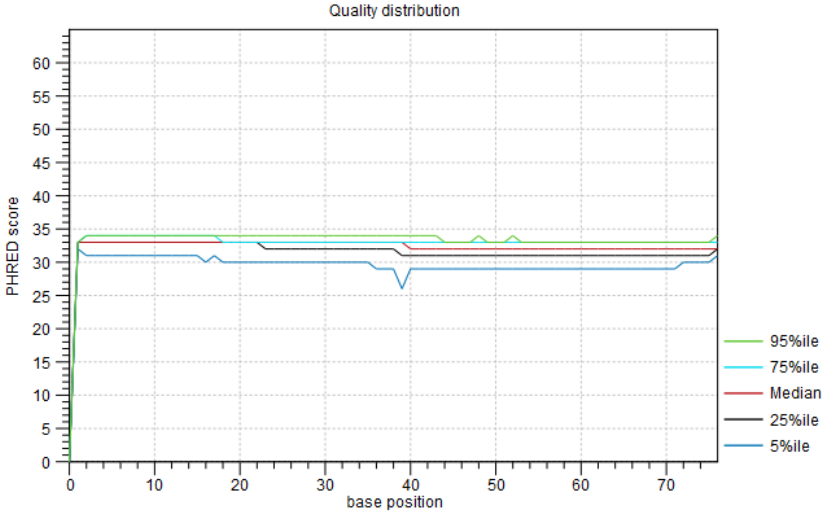
3.5 Quality distribution



Base-quality distribution along the base positions.

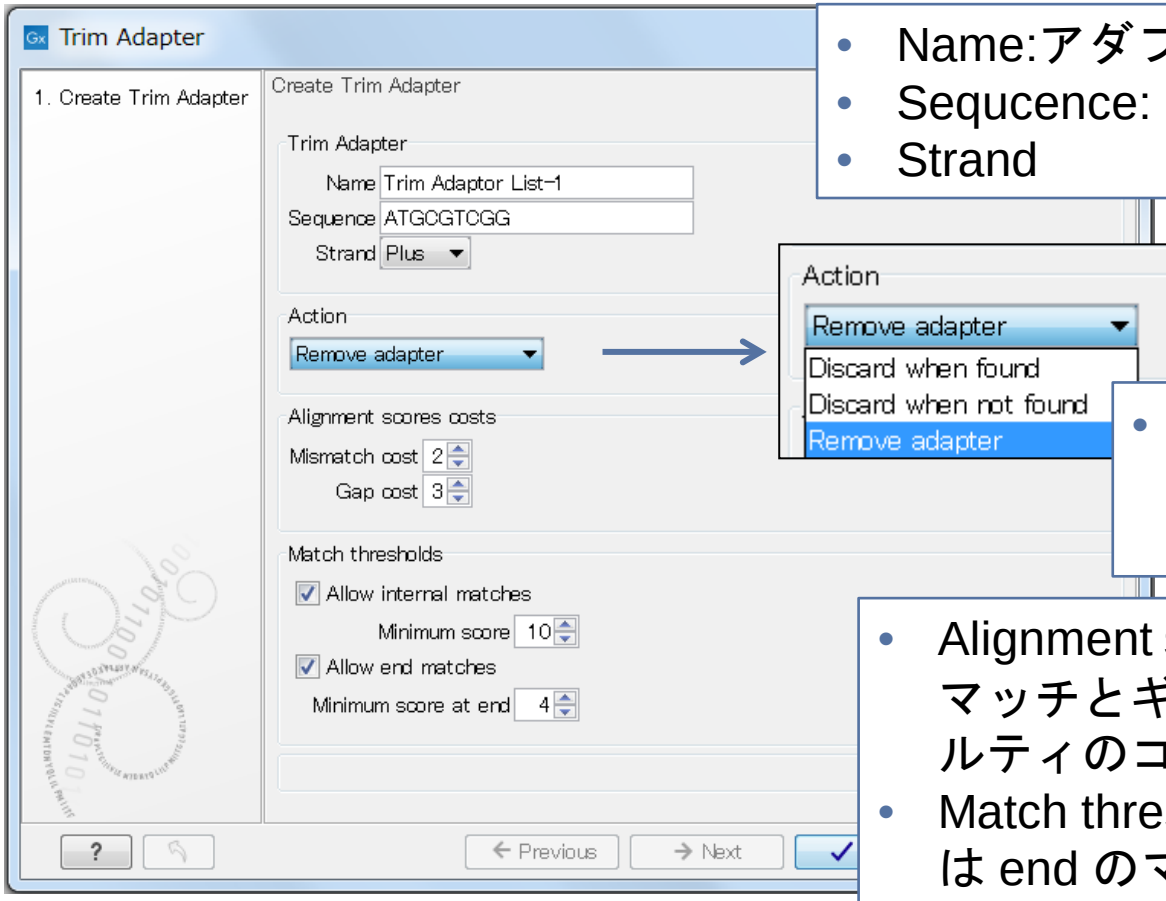
After

3.5 Quality distribution



アダプタートリミング：アダプターの指定

File > New > Trim Adapter List



- Name:アダプターの名前
- Sequcence: アダプターの配列
- Strand

- アダプターが見つかった場合のアクションの指定

- Alignment scores costs: ミスマッチとギャップに対するペナルティのコスト
- Match thresholds: internal または end のマッチに対する criteria

45

マッピング

マッピング原理

2つのステップ

1. ローカルアライメント

参照配列と似ている場所を探す

2. フィルタリング

どの程度参照配列と一致しているリードをその後の解析に残すか

マッピング原理

スコアリング

最適なマップ場所をLocal Alignmentで探索

Match = 1, Mismatch cost = 2

リード配列（20bp）が全て一致した場合

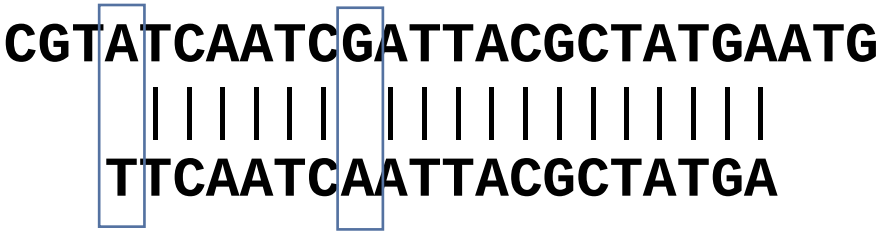
```
CGTATCAATCGATTACGCTATGAATG
|||||||||||||||||||||
ATCAATCGATTACGCTATGA
```

アライメントスコア = 20

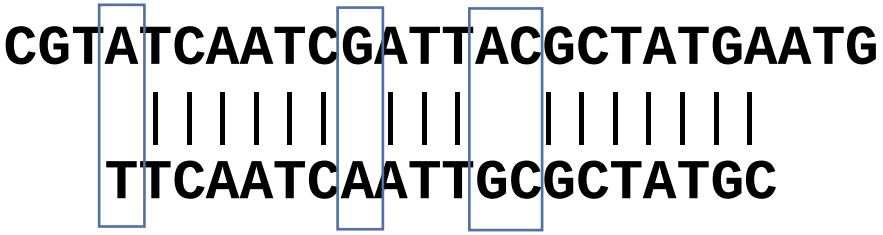
マッピング原理

スコアリング

アライメントスコア = 19



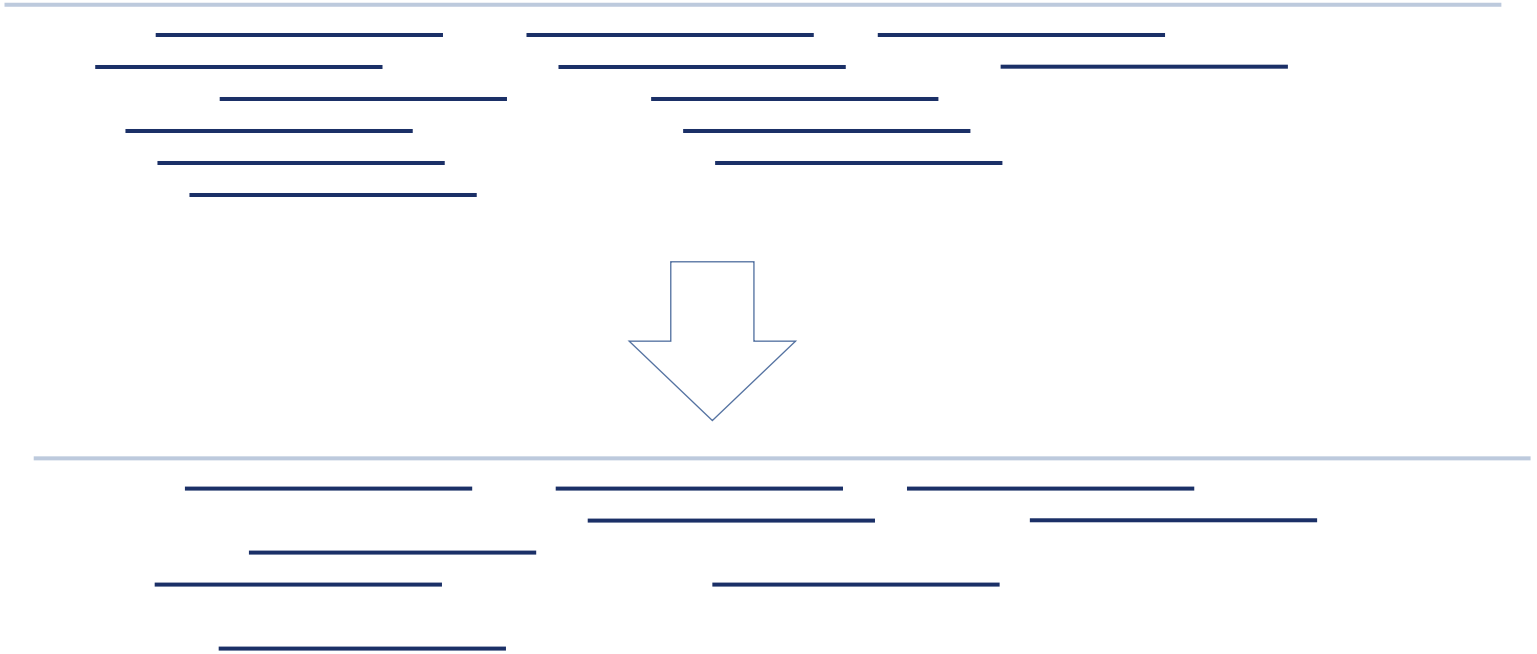
アライメントスコア = 16



アライメントスコア = 10

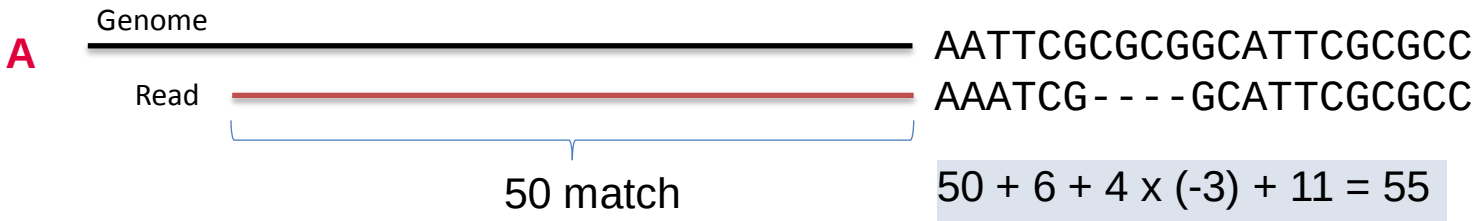
フィルタリング

最も高いアライメントスコアにマップされたリードのうち、どの程度参照配列と類似しているリードをその後の解析に残すのかを決定します。



Linear gap とAffine gap

Linear gap cost の場合（Deletionコストが３の場合）



Affine Gap costを使った場合　（Gap open = 6, Gap extend = 1）

これまでのマッピングでは、Aのように本来マッピングすべきような場合でも、リードの末端部分をアライメントしない（Bのブルーの箇所）場合のほうが、アライメントスコアが高くなるため、大きな挿入や欠失がうまくマップできていないことがありました。アフィンGapコストの場合、このような問題を防ぐことができます。またGapを開くときのコスト（Open）と延長するときのコスト（Extend）が別に設定できることで、より細かくコントロールが可能になる場合があります。

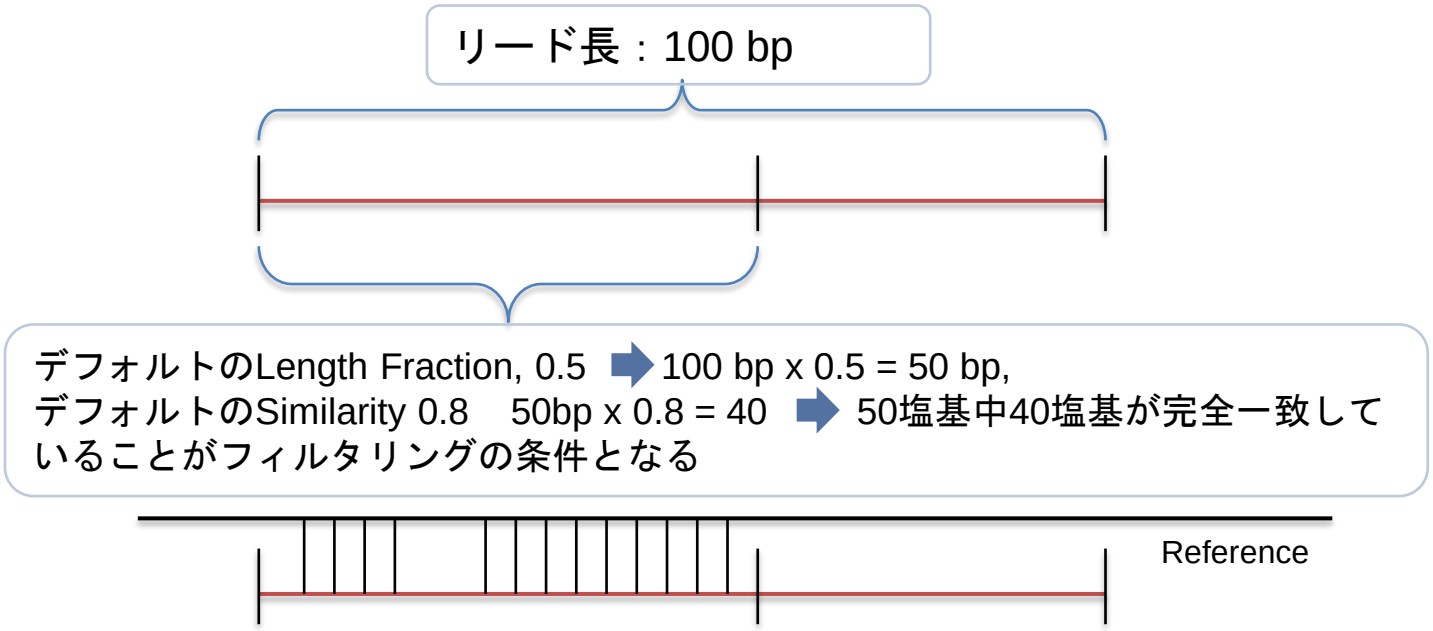
フィルタリング原理

Length FractionとSimilarity パラメータを使って、どの程度アライメントされたリードを、マッピングされたものとして保持するか、決定します。

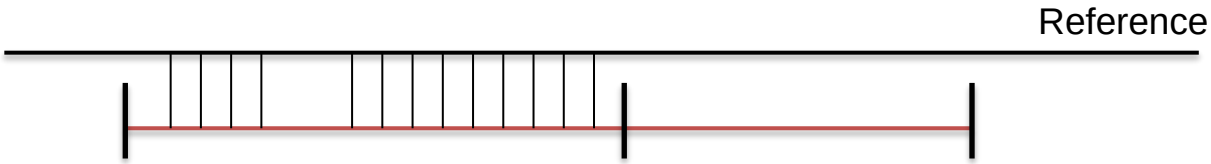
Length Fraction とSimilarity は２つのパラメータの組み合わせで使用されます。

Length fraction: フィルターをかける際に、考慮する長さ

Similarity: Length Fraction で指定した長さのうち、どの程度類似しているものを残すか。

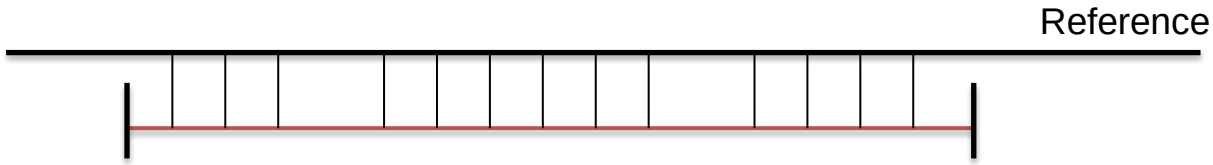


２つのパラメータを使う理由



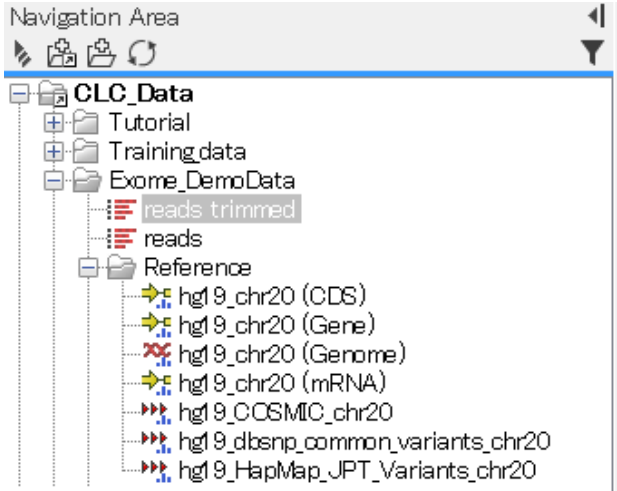
- リードの一部は似ているけれども、大きな挿入や、欠失によりリードの一部が参照配列と一致しない可能性がある場合
- トリミングが完全にできなかったクオリティの低い配列が末端部にある場合

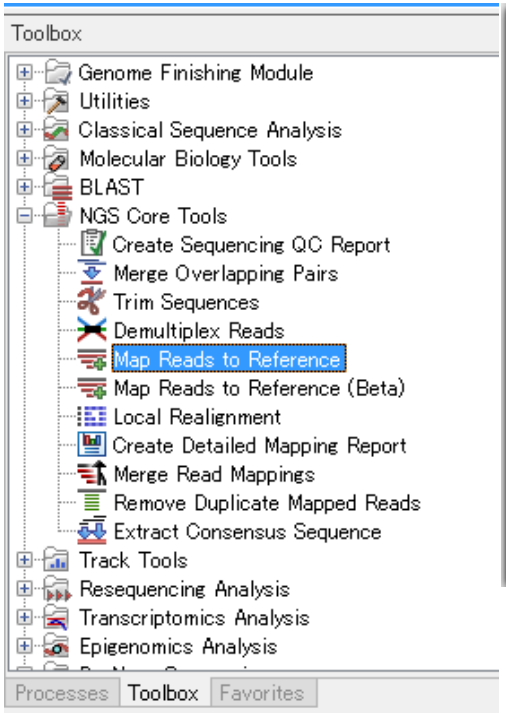
（Length Fraction を小さくすることで、リードの一部に限定してアライメントの類似度を設定できる）

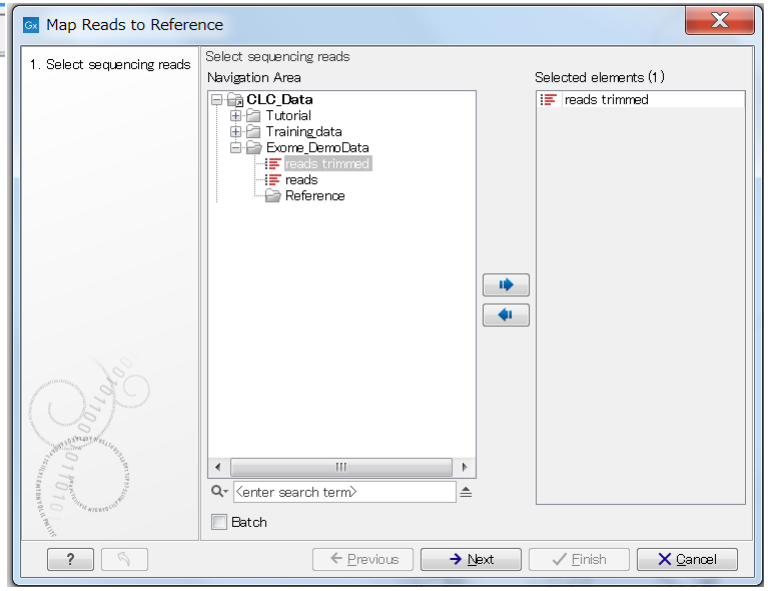


- 参照配列とほぼ一致するが、所々、１塩基の変異があると想定される場合

マッピング

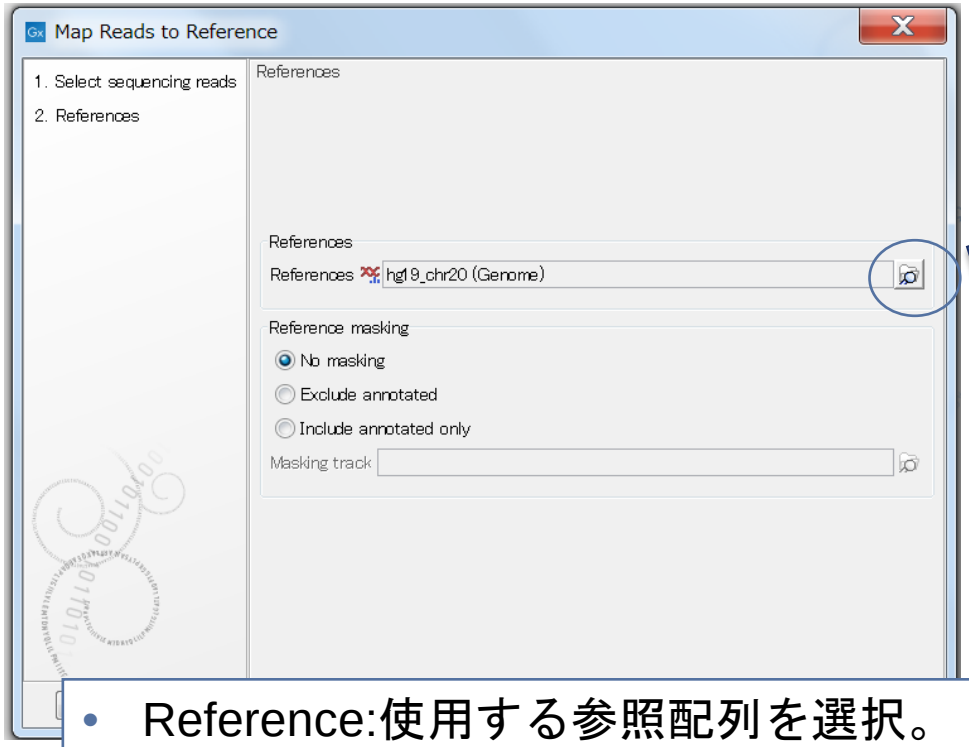


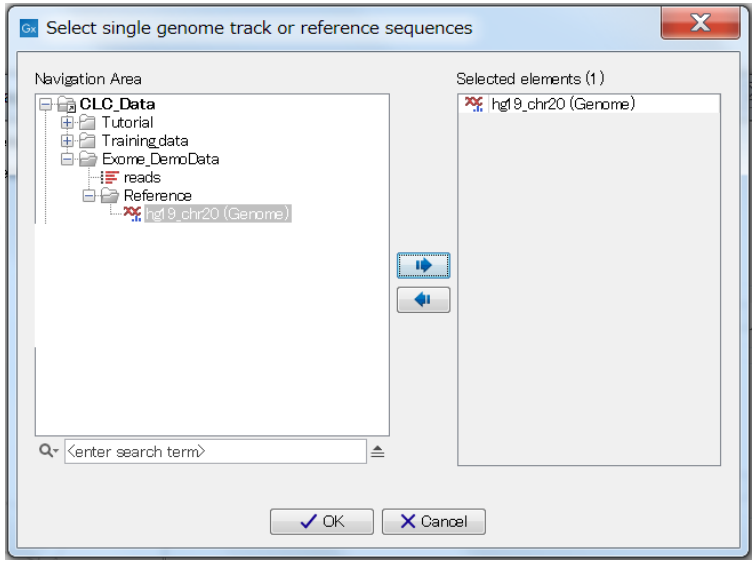




- Navigation Areaから使用するデータを選択。
- Toolboxから NGS Core Tools > Map Reads to Reference を選択、ダブルクリック。
- ウィザードが起動し、選択したデータが選ばれていることを確認。

マッピング

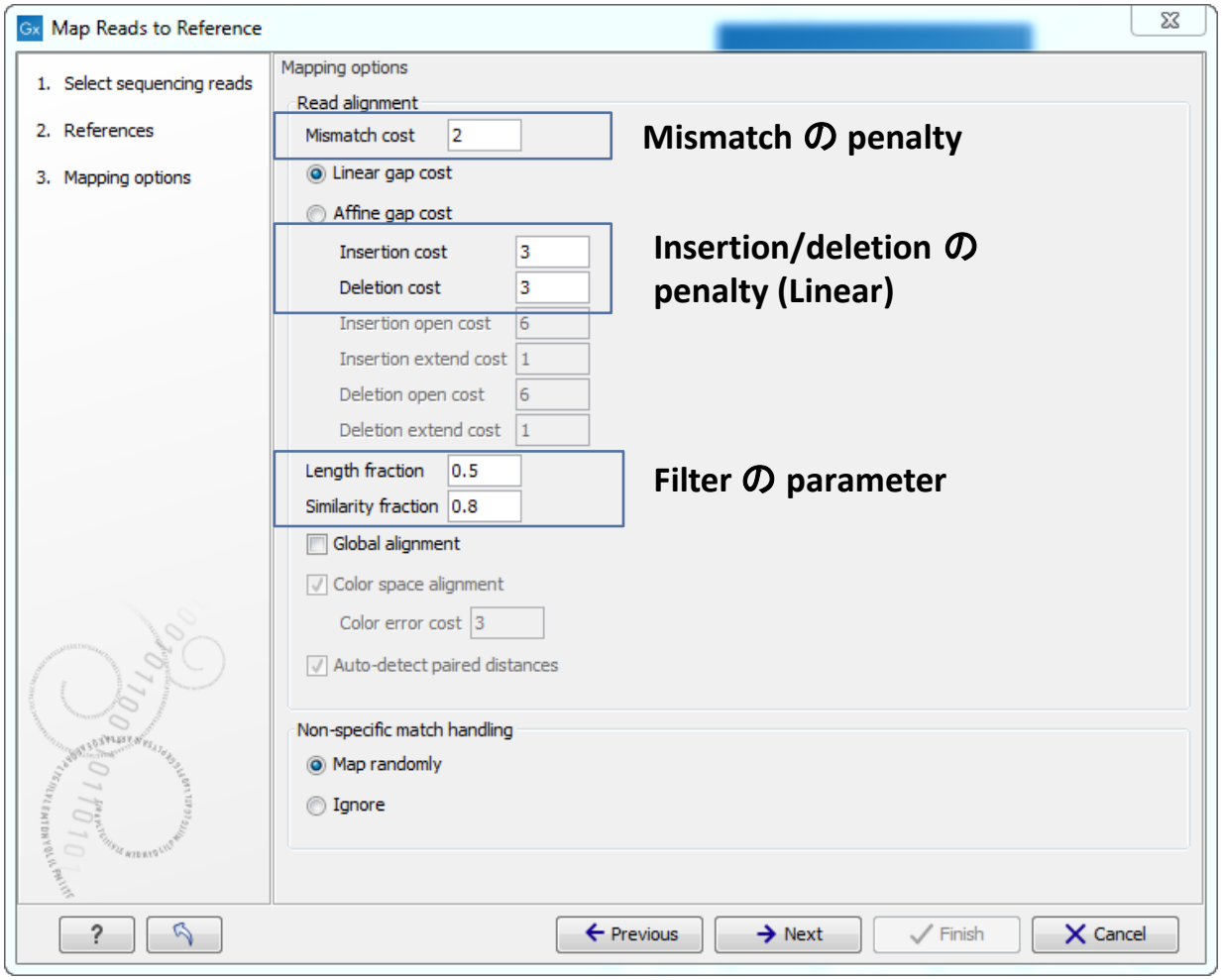




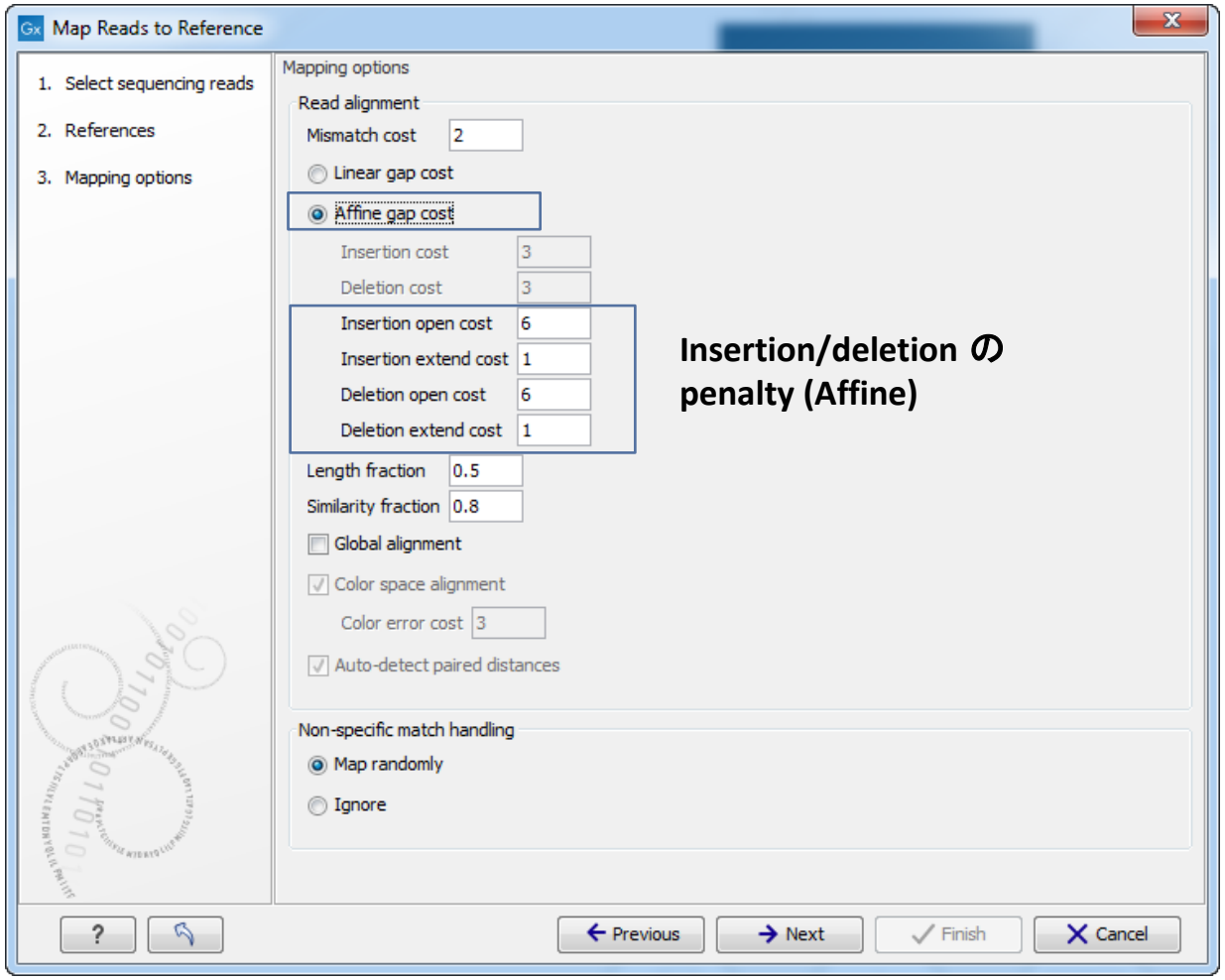
- Reference:使用する参照配列を選択。
- Reference masking
  Exclude annotated：あるアノテーションを除外したい場合。
  Include annotated only: あるアノテーションのみ含みたい場合。

- Referenceに使用するデータを選択。

Mapping parameters: Linear Gap cost



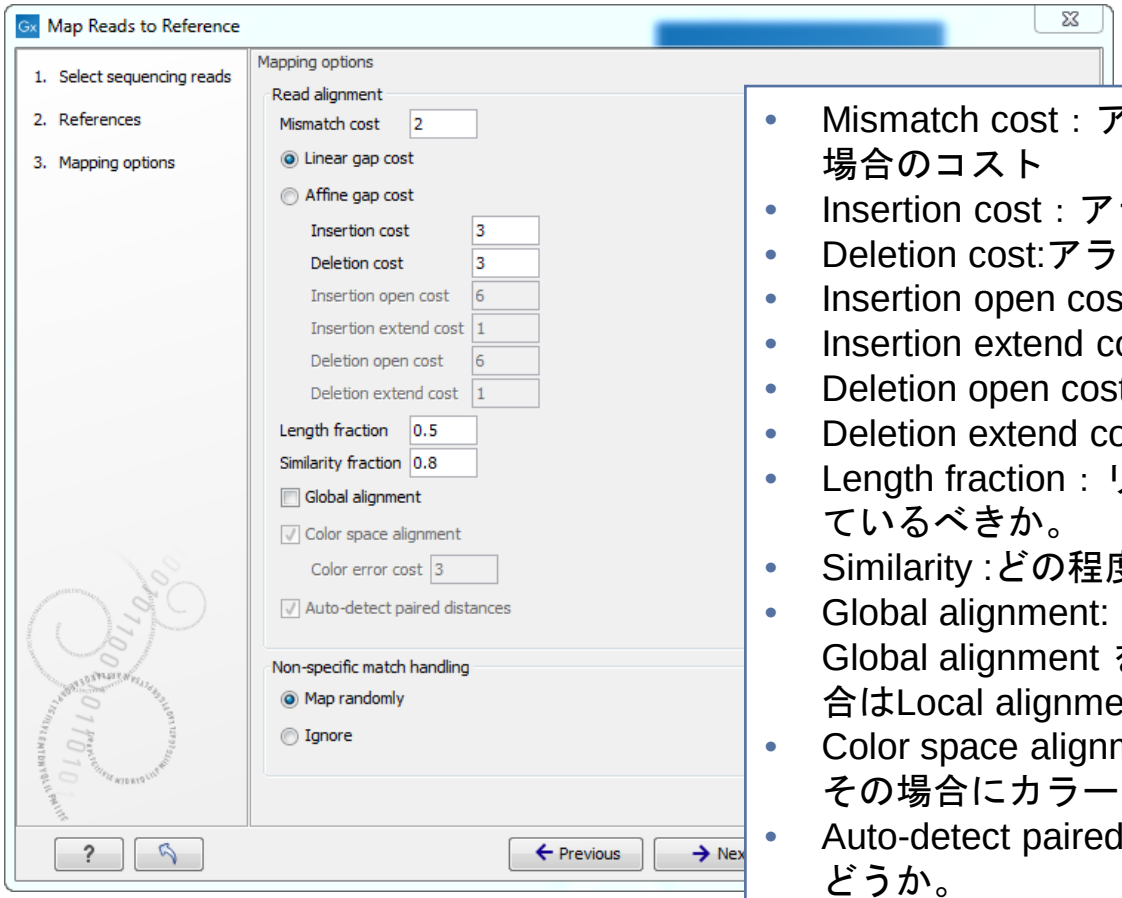
Mapping parameters: Affine gap cost



Insertion/deletion の penalty (Affine)

Insertion, Deletion 開始時点でカウントされるコスト（Insertion/Deletion open cost）と
Insertion, Deletionが1塩基長くなる際に増えるコスト（Insertion/Deletion extended cost）

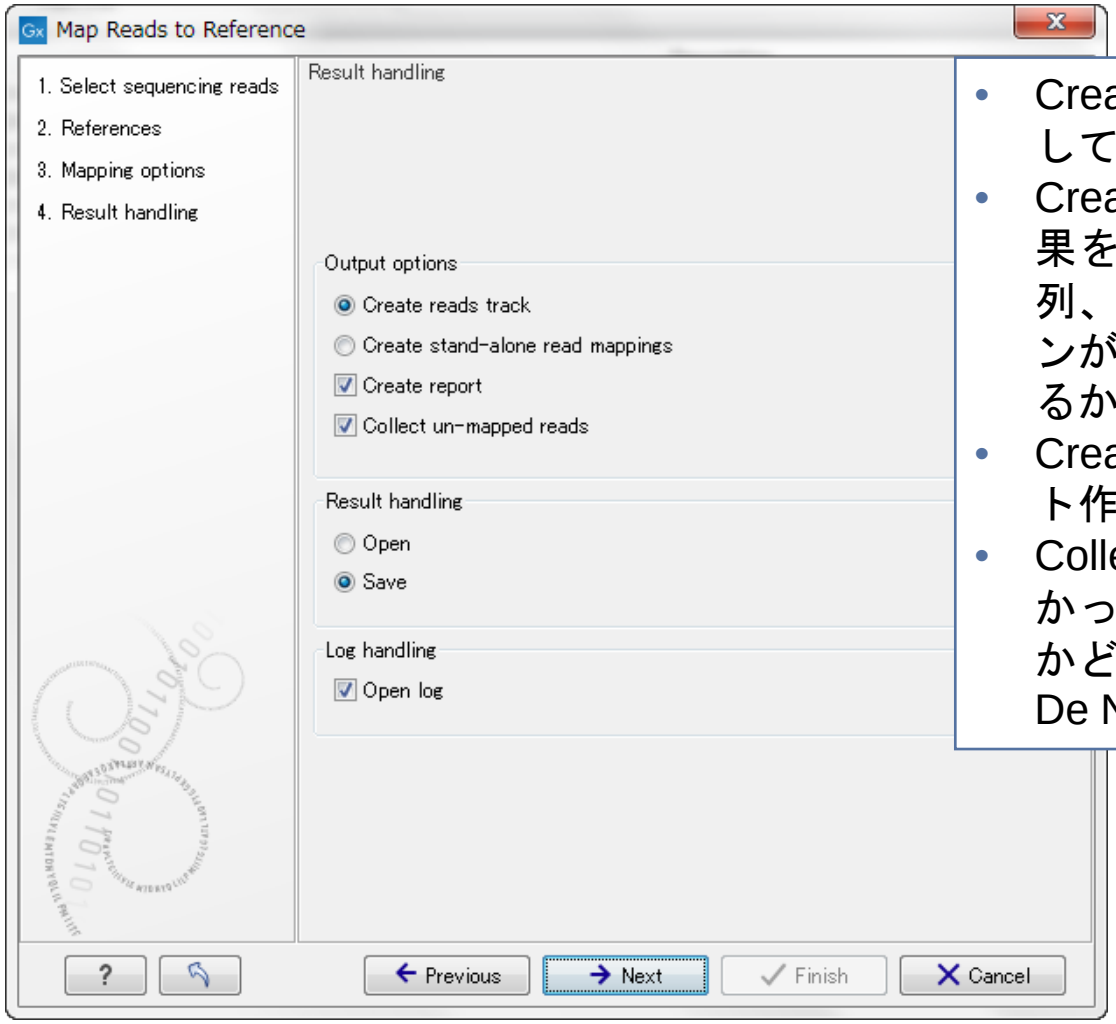
マッピング



- Mismatch cost：アライメントにマッチしないものがあった場合のコスト
- Insertion cost：アライメントに挿入がある場合のコスト
- Deletion cost:アライメントに欠失がある場合のコスト
- Insertion open cost: 挿入を開始する場合のコスト
- Insertion extend cost: 挿入を延長する場合のコスト
- Deletion open cost: 欠失を開始する場合のコスト
- Deletion extend cost: 欠失を延長する場合のコスト
- Length fraction：リードの長さのどの程度がマッピングされているべきか。
- Similarity :どの程度類似しているべきか。
- Global alignment:
  Global alignment を行うかどうか。チェックが外れている場合はLocal alignmentを実行。
- Color space alignment:カラースペースのデータかどうか、その場合にカラーによるエラー補正を行うかどうか。
- Auto-detect paired distances: 自動でペアの距離を決めるかどうか。

- Non-specific match handling：同一スコアでマップされる箇所がある場合の対処。

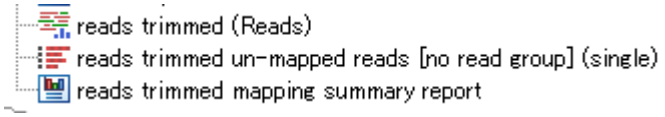
マッピング

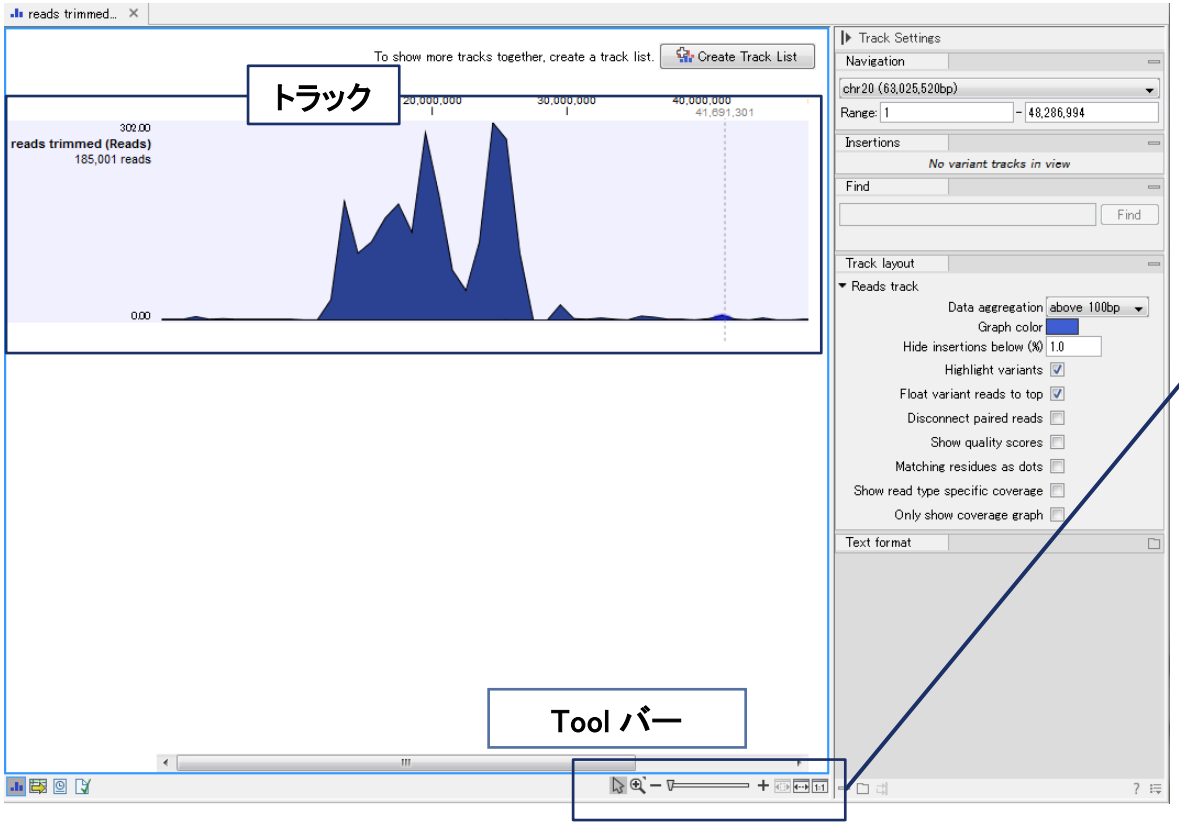


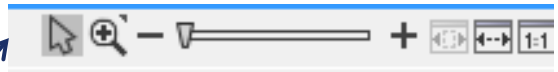
- Create reads track: 結果をトラックとして作成する場合。
- Create stand-alone read mappings：結果をstand-aloneフォーマット（参照配列、リードマッピング、アノテーションが一つになったファイル）で作成するか。
- Create report:マッピング結果のレポート作成。
- Collect un-mapped reads:マップされなかったリードをリストとして作成するかどうか（リスト化することにより、De Novoなど、別の解析へ利用可能）

マッピング：結果

結果のトラック







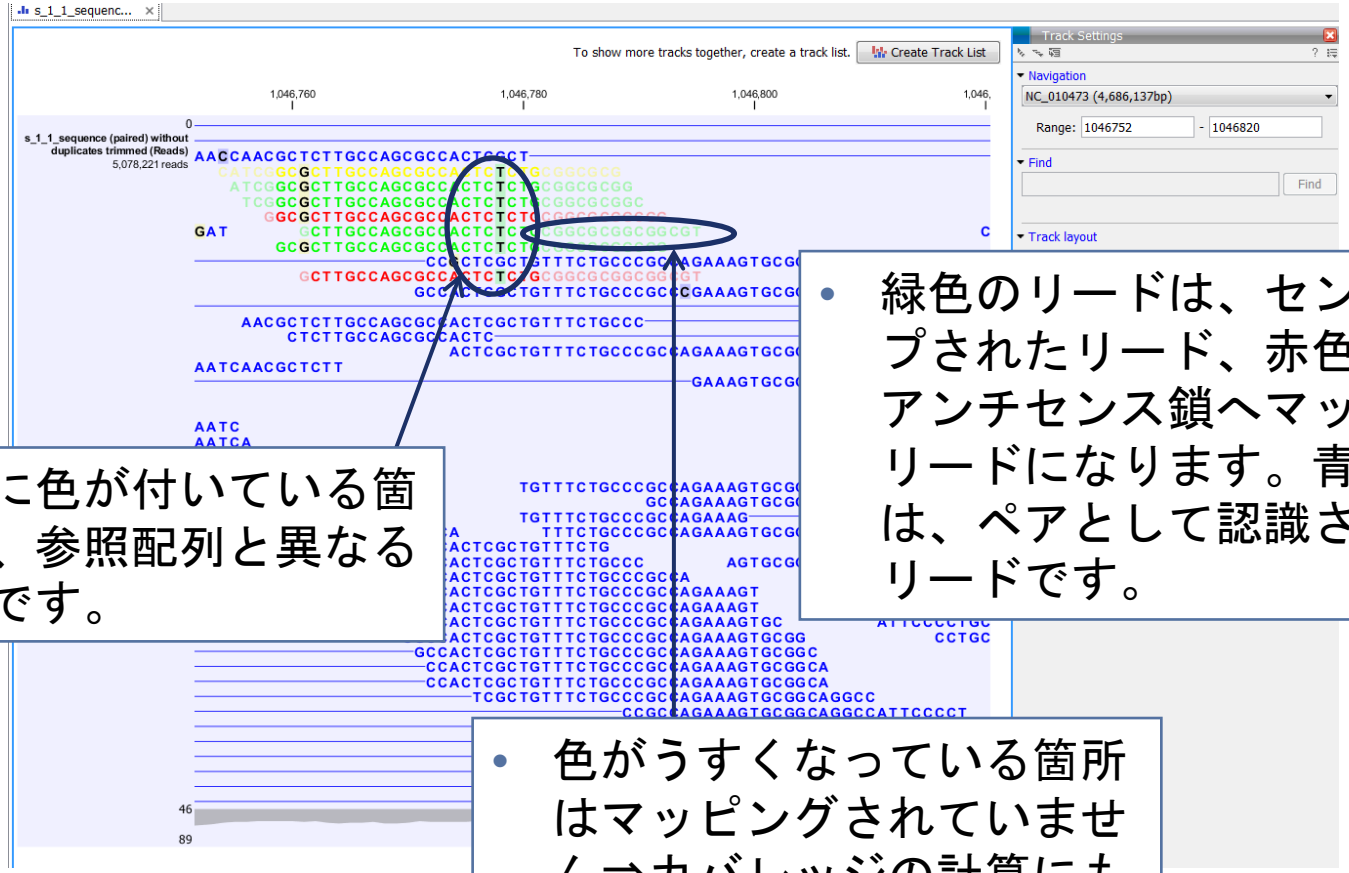
選択ツール

拡大ツール

スライドズーム

縮小して全体表示ボタン

マッピング：結果



- 緑色のリードは、センス鎖にマップされたリード、赤色のリードはアンチセンス鎖へマップされたリードになります。青色のリードは、ペアとして認識されているリードです。

- 背景に色が付いている箇所は、参照配列と異なる箇所です。

- 色がうすくなっている箇所はマッピングされていません⇒カバレッジの計算にも考慮されていません。

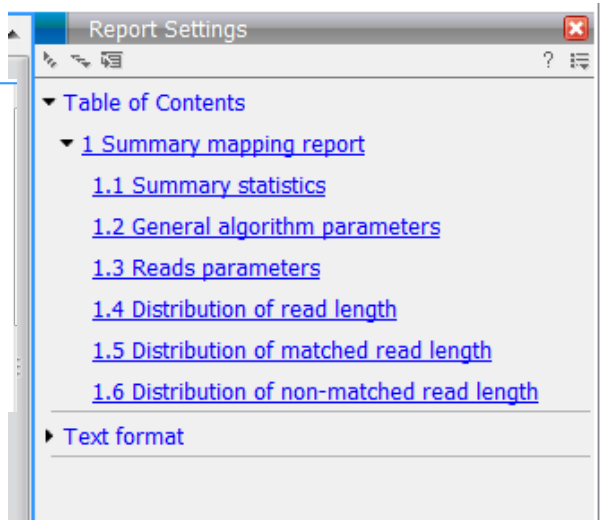
マッピング：レポート

基本の Report は「Summary Report」という名前で保存されています。

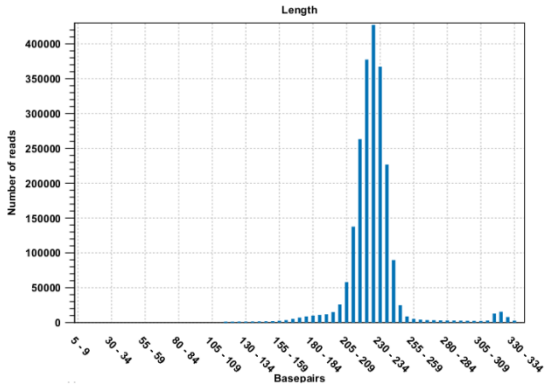
1 Summary mapping report

1.1 Summary statistics

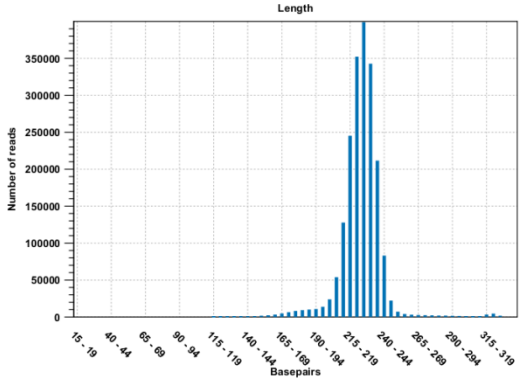
	Count	Percentage of reads	Average length	Number of bases	Percentage of bases
References	1	-	4,686,137.00	4,686,137	-
Mapped reads	5,078,221	97.95%	34.00	172,651,355	97.92%
Not mapped reads	106,167	2.05%	34.62	3,675,884	2.08%
Reads in pairs	4,987,262	96.20%	215.38	170,198,371	96.52%
Broken paired reads	90,959	1.75%	26.97	2,452,984	1.39%
Total reads	5,184,388	100.00%	34.01	176,327,239	100.00%



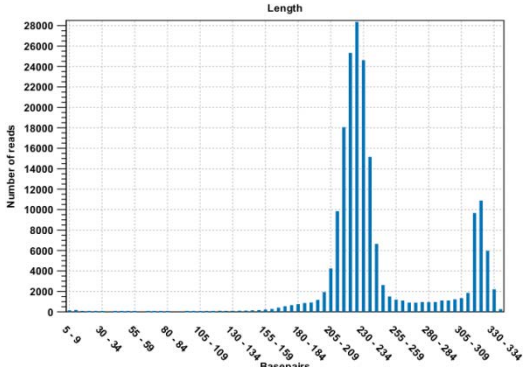
1.4 Distribution of read length



1.5 Distribution of matched read length



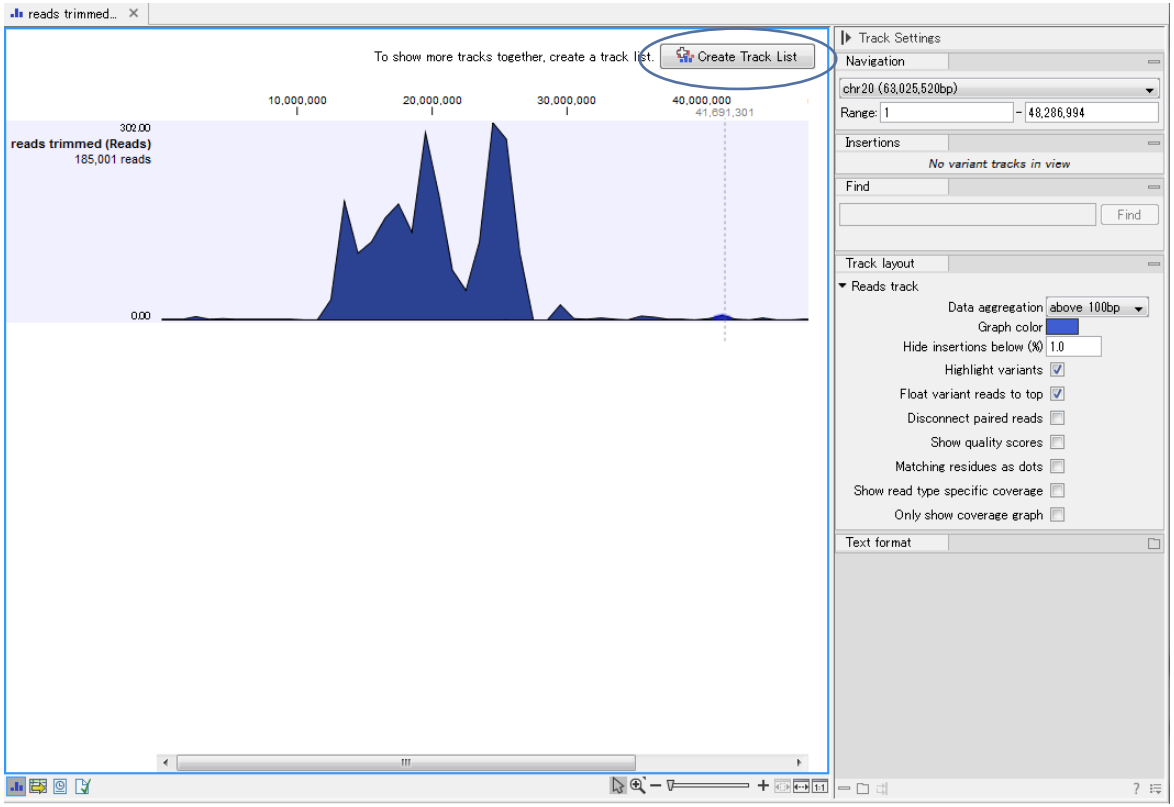
1.6 Distribution of non-matched read length



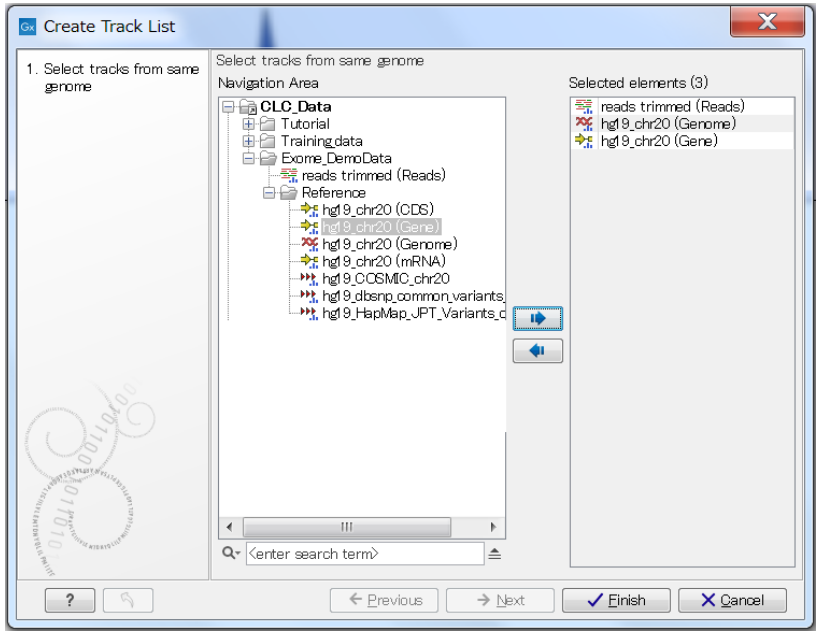
マッピング：Track list の作成

参照配列の追加

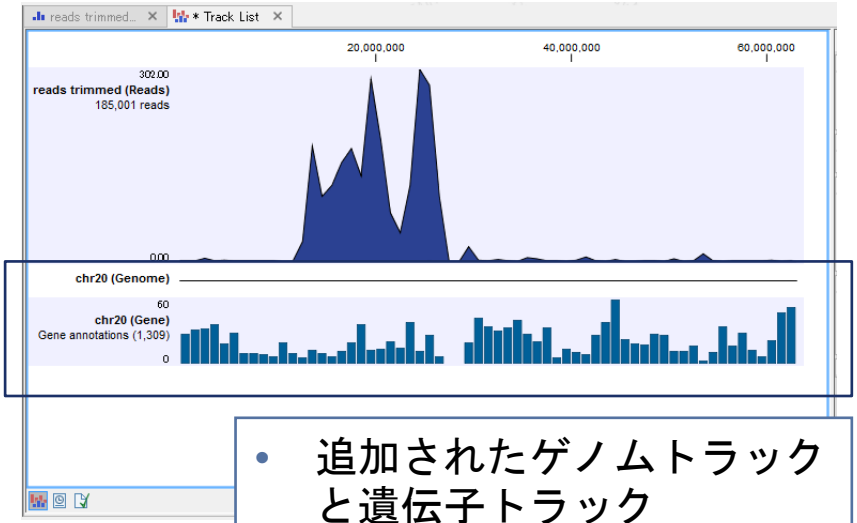
リードマッピングの結果に参照配列を追加しましょう。



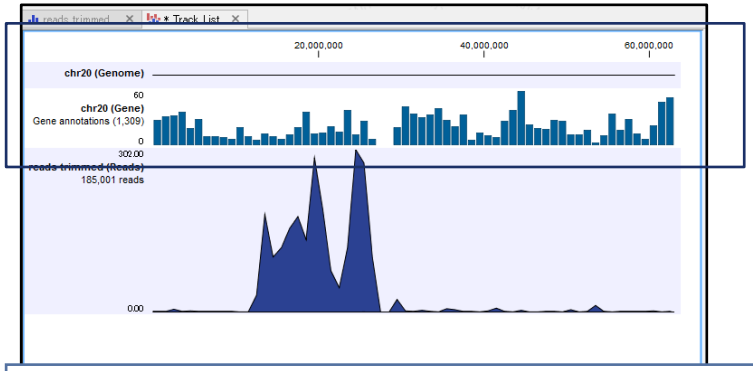
マッピング：Track list の作成



- マッピングに使用したゲノムを選択。
- ゲノム（参照配列）のアイコンが の ような場合、Track Tools > Convert to Track を使って、変換を行ってください。



- 追加されたゲノムトラックと遺伝子トラック



- ドラッグアンドドロップで簡単に位置を変更できます。

Local Realignment

原理

マッピングのプロセスでは、各リードがもっとも高いアライメントスコア（参照配列との一致度を示すスコア）を示す場所にマッピングをしています。しかしながら、時には近傍のリードのマッピングの状況から、最も高いアライメントスコアではなくとも、もっともらしいマッピング結果が考えられる場合があります。

たとえば上記例では、GCCGは左横にずれることで、他のリードのマッピングとも一致しもっともらしいマッピングになると考えられます。マッピングの段階では、各々のリードのアライメントスコアのみを考えているため、このような状況が発生します。

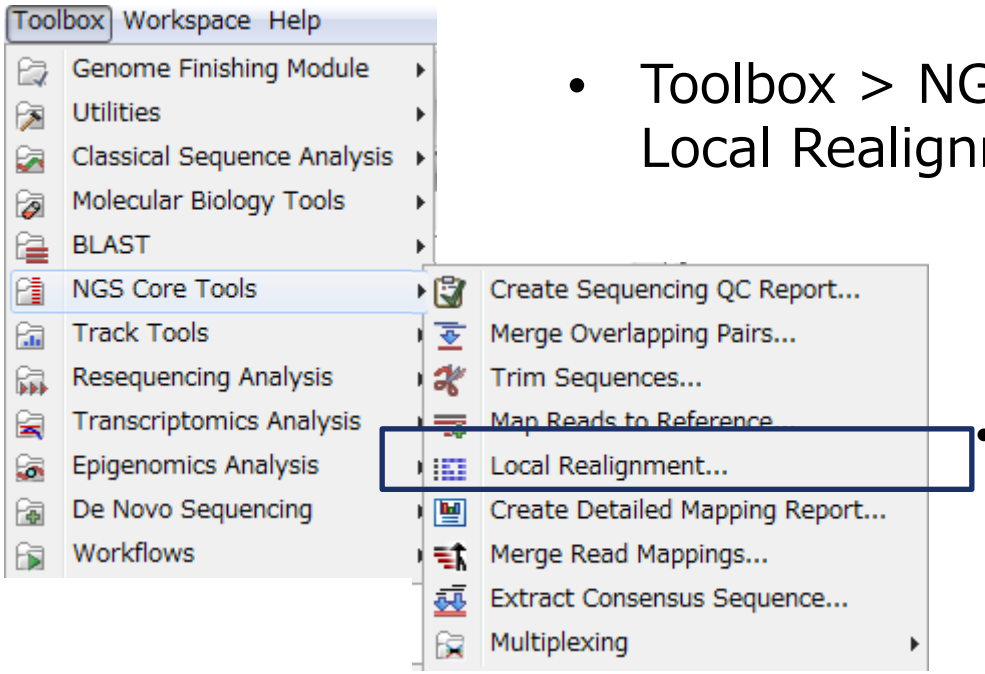
さらにこの状況で変異やInsertion、Deletionの検出を行うと、正しく検出できないものも発生します。特にInsertionやDeletionが影響をうけると考えられています。

原理

- Local Realignment では、このような状況を修正するため、マッピングを部分的にやり直します。この際、通常のマッピングの段階とは異なり、他のリードのマッピング状況を考慮するため、先ほどのマッピングは以下のように変化します。

- 先ほどのマッピングよりも、こちらの方がもっともらしい結果であることが直感的に分かります。

Local Realignment



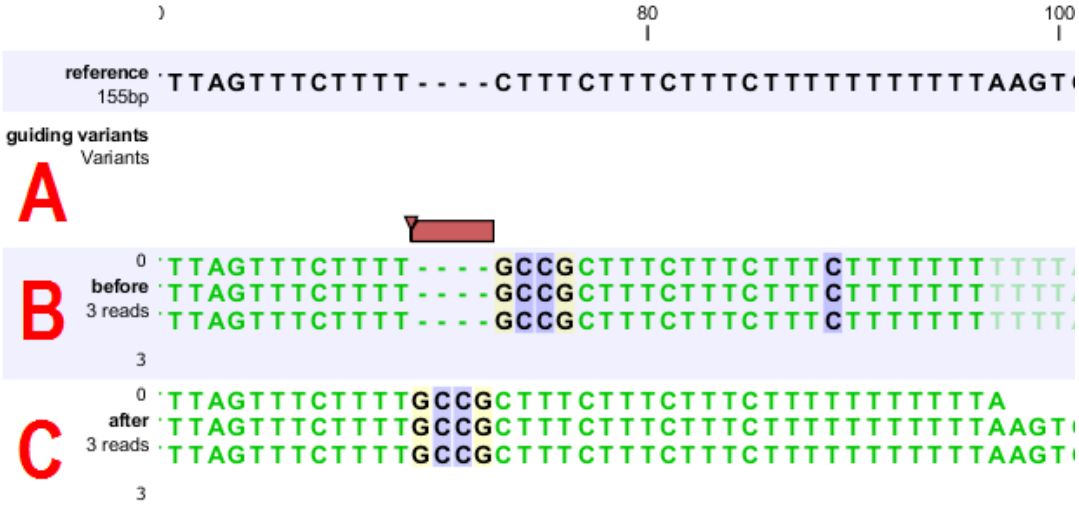
- Toolbox > NGS Core Tools > Local Realignment

- 2種類のLocal Realignmentsがあります。さらにGuided にNo forceとForceの2種類があります。
  - Non guided
  - Guided
    - No force
    - Force

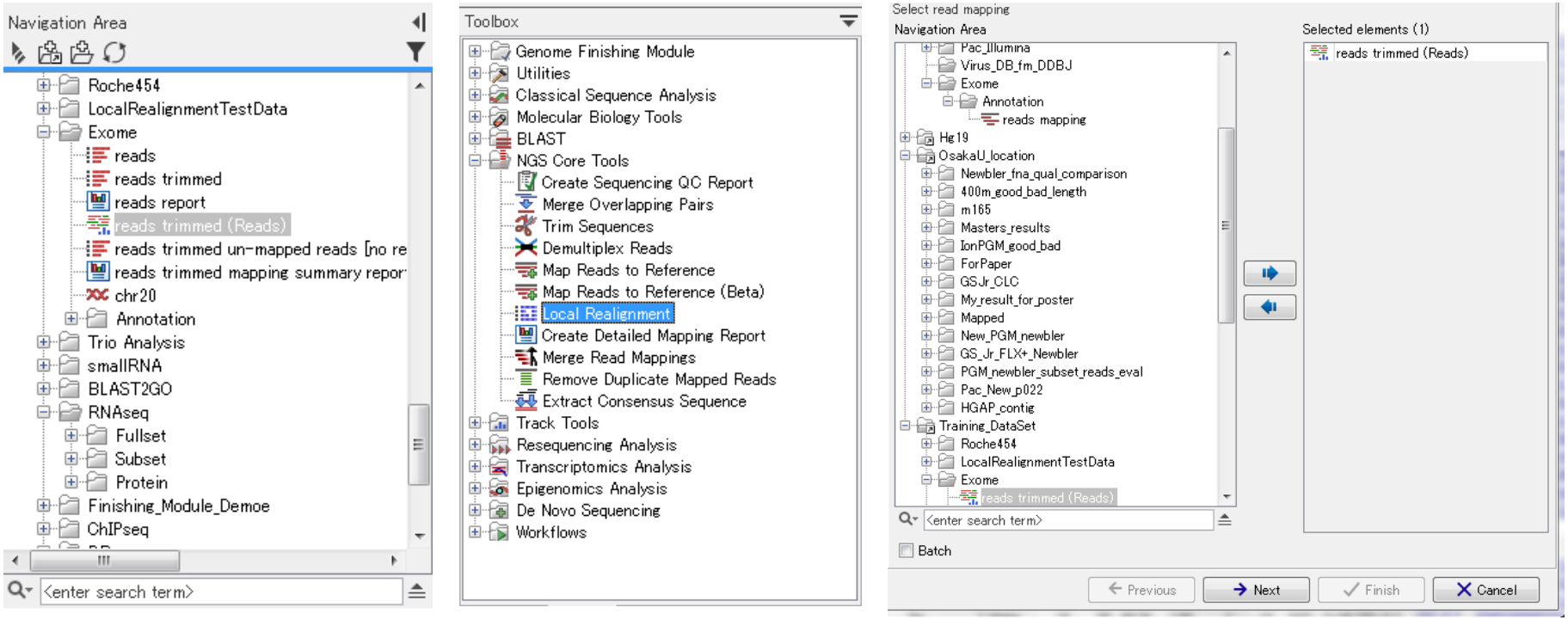
Guided Local Realignment

- ガイドとなるような変異（InsertionやDeletion）の情報をあらかじめ与えておくことで、その領域のInsertion、Deletionを考慮してリアライメントを行う。
- ガイドとなる変異情報がない場合、Local Realignment では、少なくとも1本のリードがInsertionやDeletionを支持している必要がある。このような場合、ガイドとなる変異情報を与えることで、InsertionやDeletion を効率的に検出できるようになる。

- Guided Local Realignment が有効な例

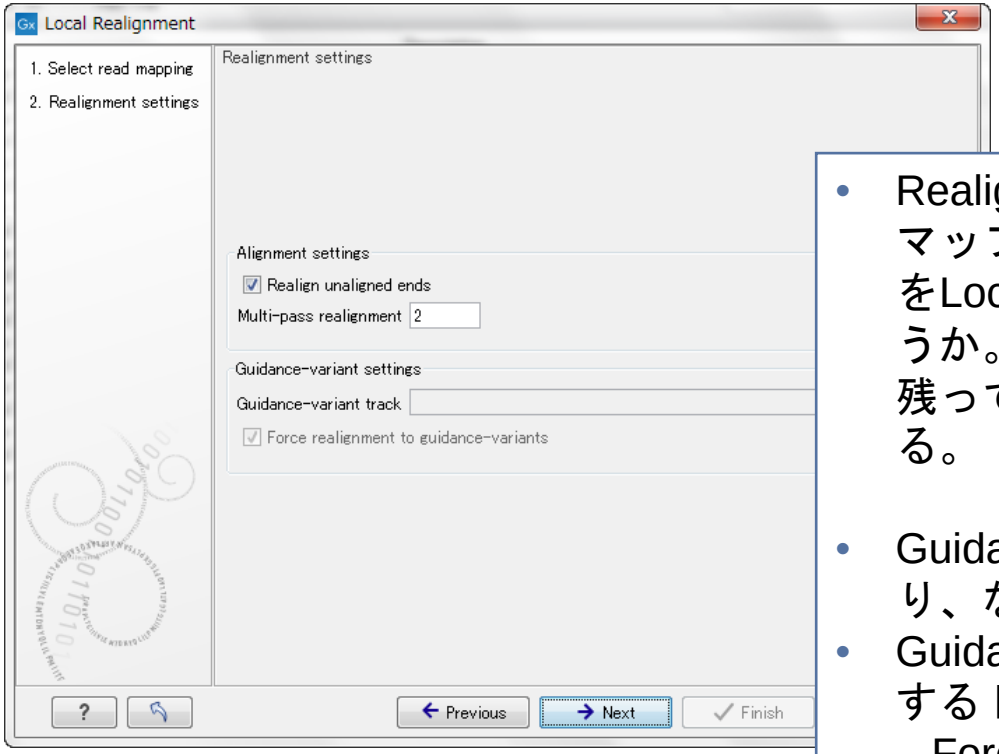


Local Realignment 実行方法



- Navigation Areaから使用するマッピングデータを選択。
- Toolboxから NGS Core Tools > Local Realignment 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。

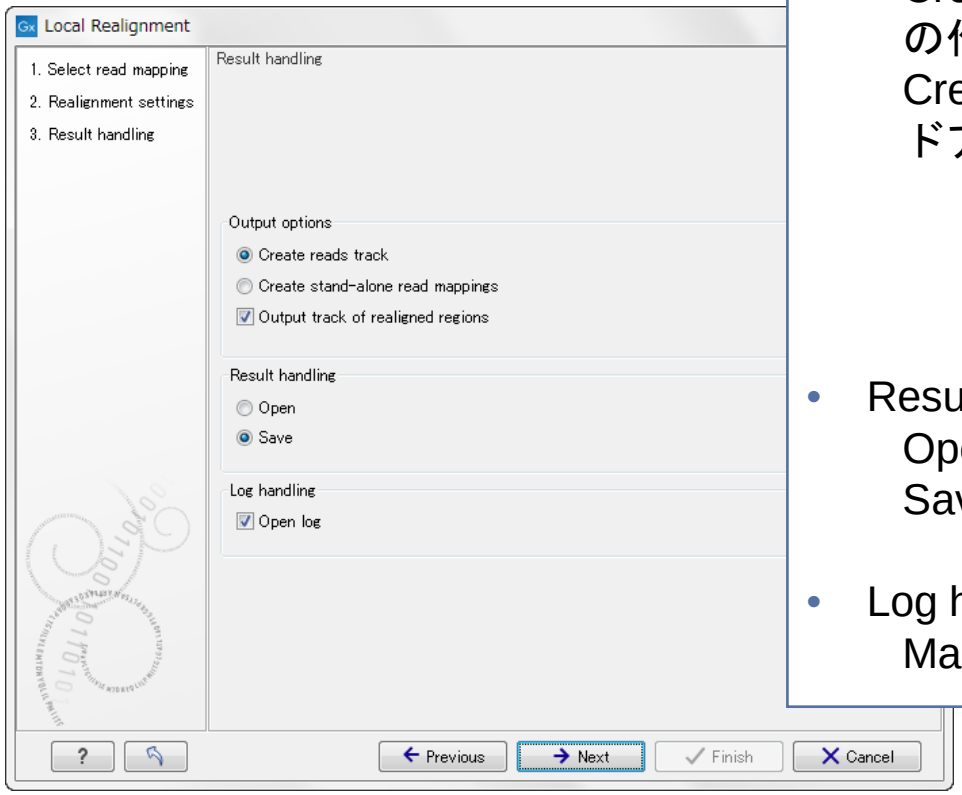
Local Realignment 実行方法



- Realign unaligned ends: マッピングの際にマップされなかった末端（soft clipping）をLocal Realignment の際に利用するかどうか。アダプターの一部のようなものが残っていない限り、ここはチェックを入れる。

- Guidance-variant settings：ガイダンスあり、なしの設定
- Guidance-variant track: ガイダンスに使用するトラックを選択。
  Force realignment to guidance-variants: ここにチェックを入れることで、より積極的にRealignmentを行える。

Local Realignment 実行方法



- Output options アウトプットの選択
  Create reads track: トラックフォーマットでの作成。
  Create stand-alone read mappings:　スタンドアロンフォーマットでの作成。
    Output track of realigned regions: Realignmentされた個所をトラックとして保存するかどうか。確認に便利。

- Result handling
  Open: 実行後すぐに開く。
  Save: 実行後一旦保存。

- Log handling
  Make log: ログを作成するかどうか。

Local Realignment：結果

- 結果はマッピングのファイルとして作成され、名前の最後に locally realigned として作成されます。
  - スタンドアロンフォーマットで作成した場合

  - トラックフォーマットで作成した場合

- この後、通常と同じ方法で変異やInsertion, Deletion の検出を行います。

変異検出

SNV検出

3種類の variant detection tools

Basic Variant Detection ：クオリティと、バリアントの見られる頻度からバリアントのサイトを検出
(version 7.5以前のQuality-Based Variant Detection)

Fixed Ploidy Variant Detection：確率モデルを使い、バリアントのサイトを検出
(version 7.5以前のProbabilistic Variant Detection)

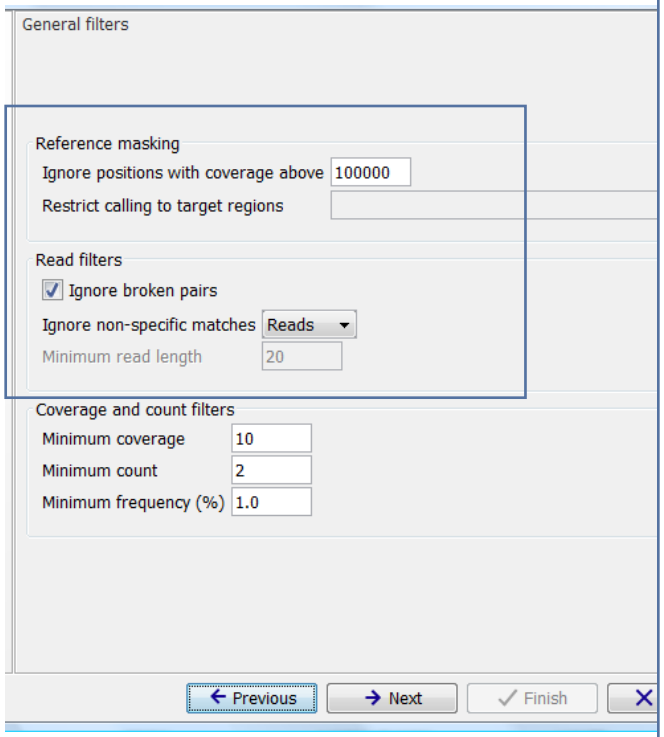
Low Frequency Variant Detection：低頻度で見られるバリアントの検出ツール。倍数性を指定しないでバリアントの検出が行える。

使い分け：

バリアントの見られる頻度が、その領域において15%以下のような場合は、Basic Variant Detection, それよりも多い場合は、Fixed Ploidy Variant Detection をご利用ください。バリアントの見られる頻度が低い場合や、倍数性を指定できない場合などは、Low Frequency Variant Detection をご利用ください。

フィルター：General Filter

共通フィルター



Reference masking

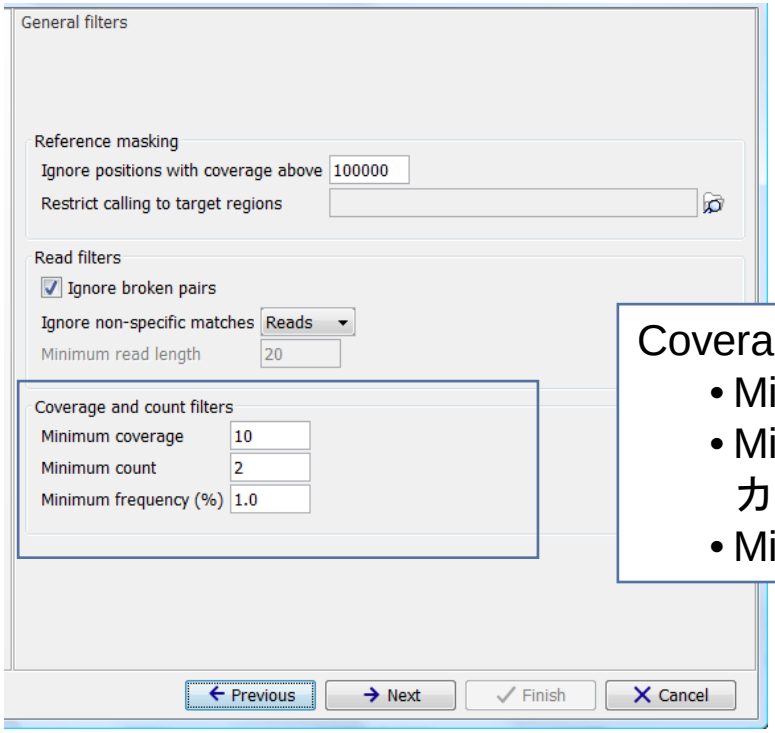
Ignore positions with coverage above：カバレッジが指定した数字以上のバリアントについてリストに含めない

Restrict calling to target regions：バリアントを検出したい領域の指定（アノテーショントラックで指定）

Read filters

- Ignore broken pairs：ペアエンドのリードでペアと認識されなかったリードをバリアント検出の計算に含めるかどうか
- Ignore non-specific matches：「Reads」を選択すると、non-specificなマッチのリードを計算に含めなくなり、「Regions」を選択すると、1本でもnon-specificなリードが含まれる場合、その領域のバリアントを検出しません。
- Minimum read length：Ignore broken pairとIgnore non-specific regions が指定された場合、このフィルターの対象となる最小のリードの長さの設定が必要です。これは非常に短いリードは、その短さからnon-specificになる可能性があるためです。

共通フィルター

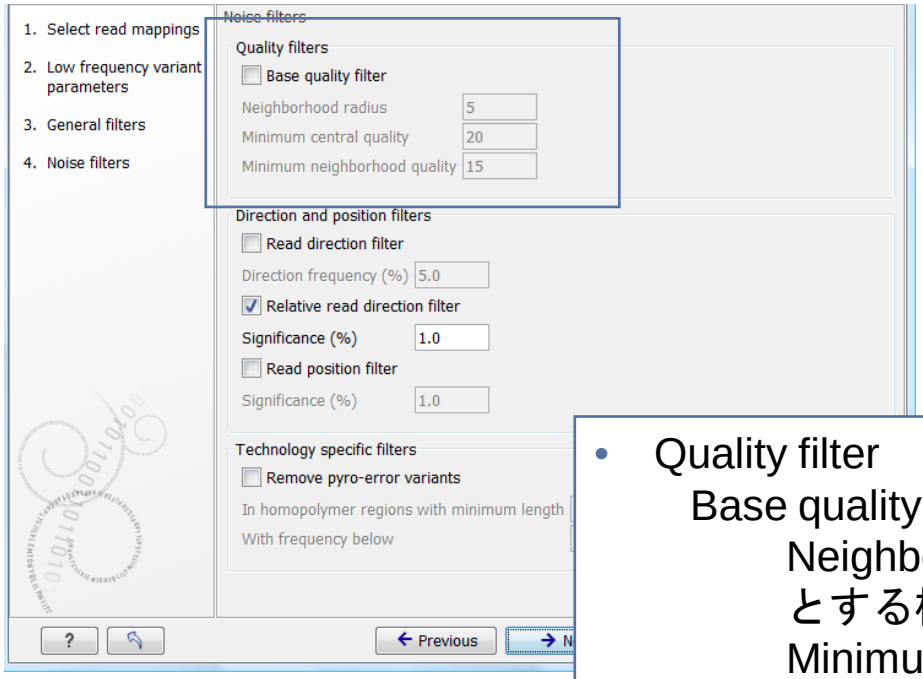


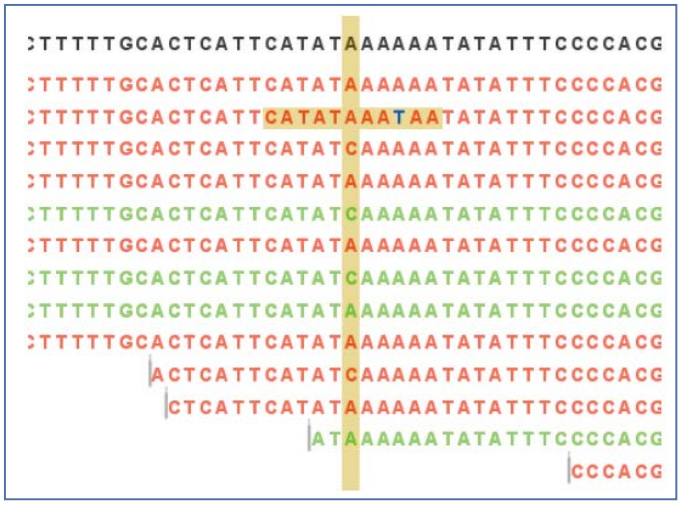
Coverage and count filters

- Minimum coverage：最小カバレッジ
- Minimum count：バリアントを支持するリードの最低カウント数
- Minimum frequency (%)：最小頻度

フィルター：Noise Filters

共通フィルター

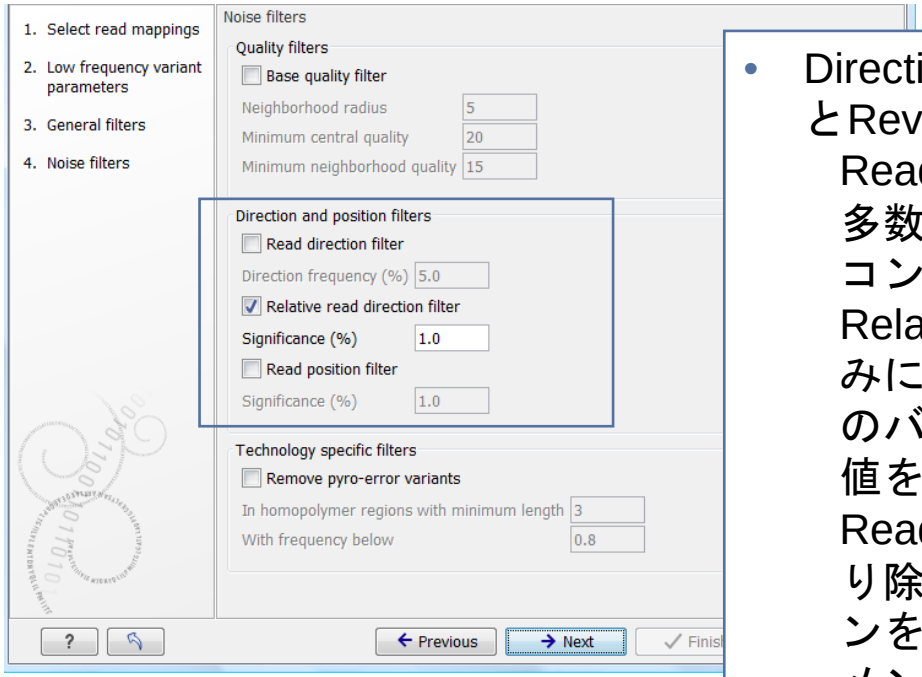




- Quality filter
  - Base quality filter:塩基のクオリティに関するフィルター
    - Neighborhood radius：クオリティフィルターの対象とする横方向の塩基数（奇数）
    - Minimum central quality：縦方向の数（リード数）
    - Minimum neighborhood quality：Neighborhood radiusで指定した範囲の最低クオリティ（Phred score）

78

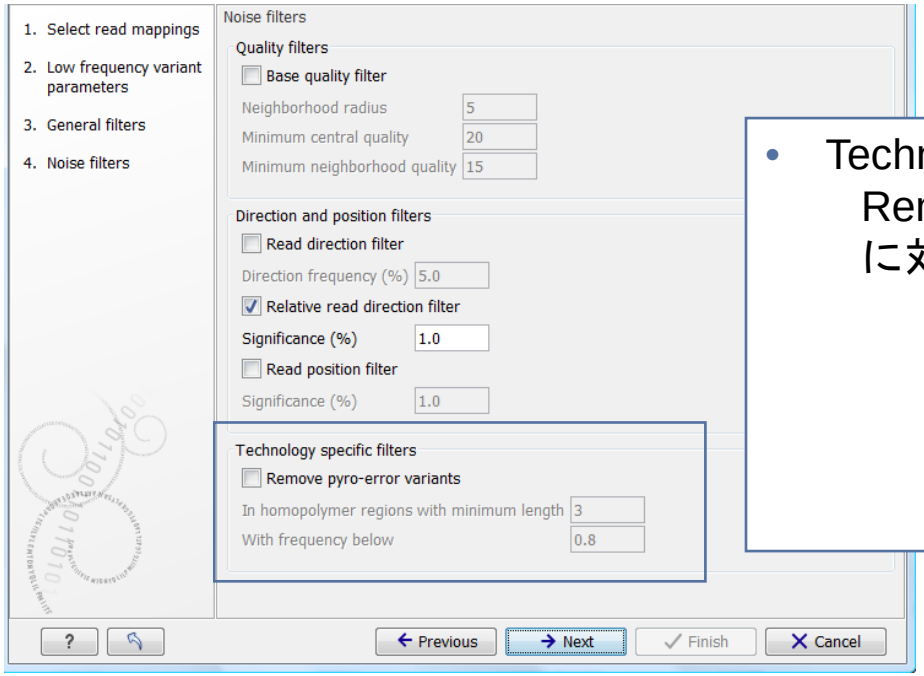
共通フィルター



- Direction and position filters：リードの方向（ForwardとReverse）とポジションを使ったフィルター
  Read direction filter：どちらか一方の方向のリードが多数見られる場合にそれを排除（ただし、アンプリコンには適していません）。
  Relative read direction filter：リードの方向が一方のみに偏りすぎていないか、全体のForwardとReverseのバランスを見て統計検定を行う。Significanceで閾値を入力。
  Read position filter：システマティックなエラーを取り除くために用いるツールでハイブリダイゼーションを行った場合のデータに有効。リードを5つのセグメントに分割し、バリアントの見られるポジションの５つのセグメントに分割されたリードの分布が全体のそれと似ているかどうか検定を行う。
  Significanceで閾値を入力。＊詳細は後述

79

共通フィルター



- Technology specific filters
  - Remove pyro-error variants：ホモポリマー領域に対するエラーの除去
    - In homopolymer regions with minimum length：指定した長さのホモポリマー領域のInDelを取り除く。
    - With frequency below：指定した頻度以下のものについてのみフィルターを適用。

Base quality filter

フィルター例

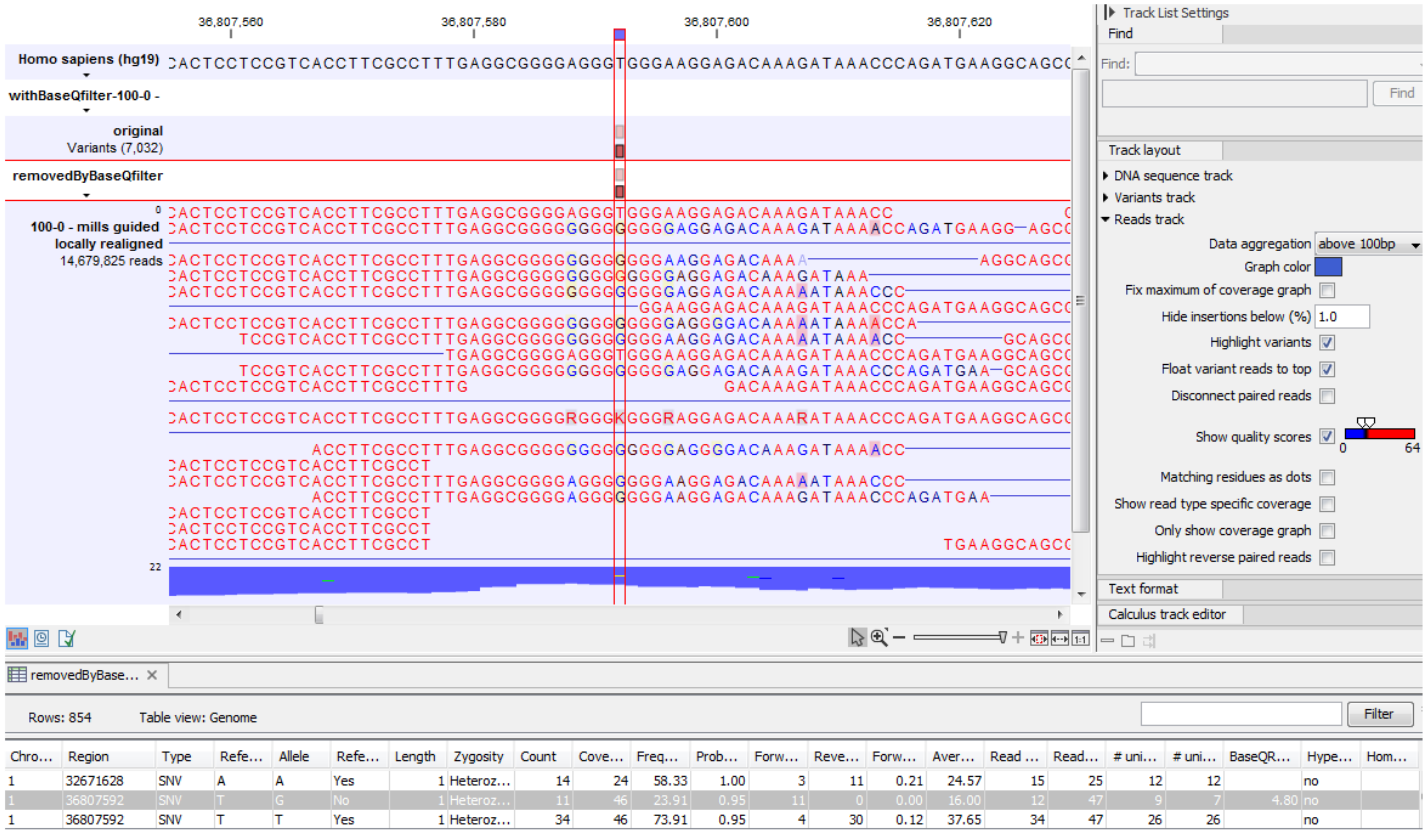
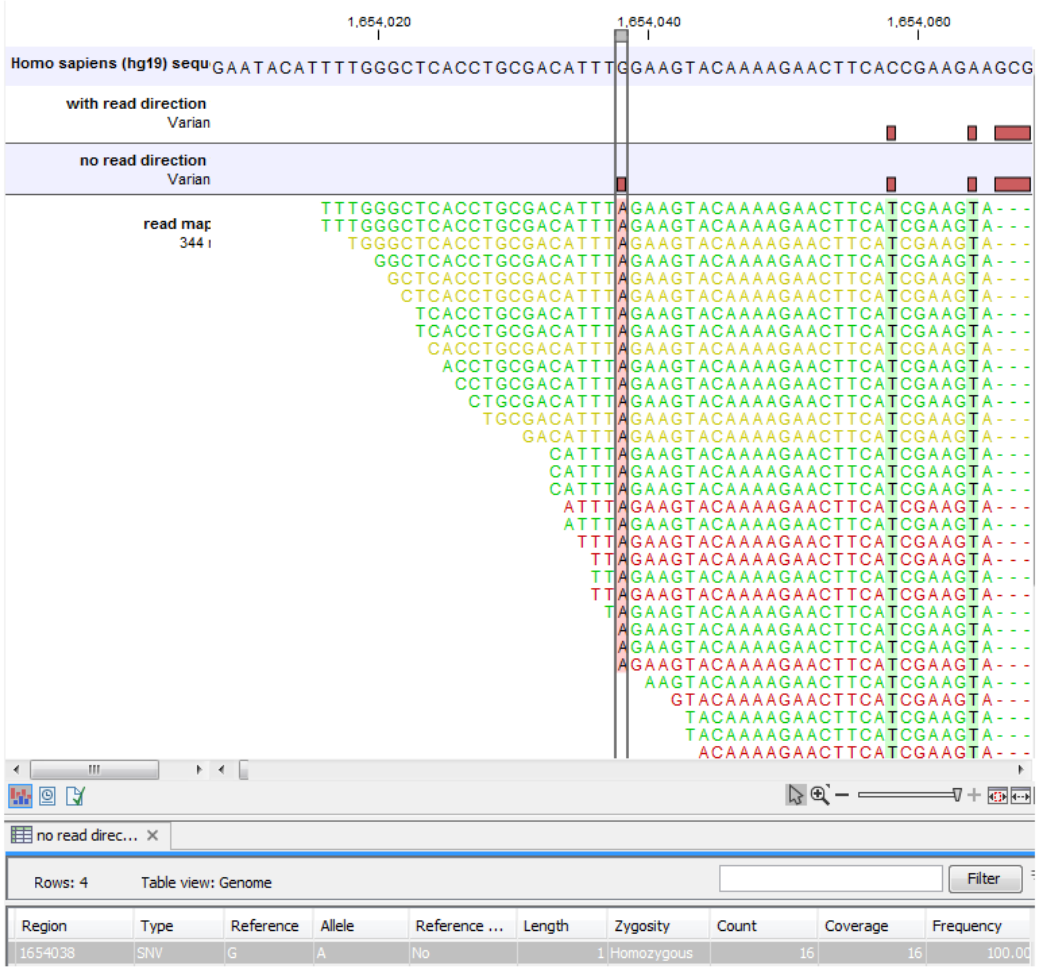


Figure 27.22: *The same data as in figure 27.21, now with the 'Show quality scores' option in the reads track switched on.*

- Base quality filter 適用例：マッピングしたリードをクオリティで表示。クオリティの低いリードがマップされている箇所がバリアントのリストからはずされます。

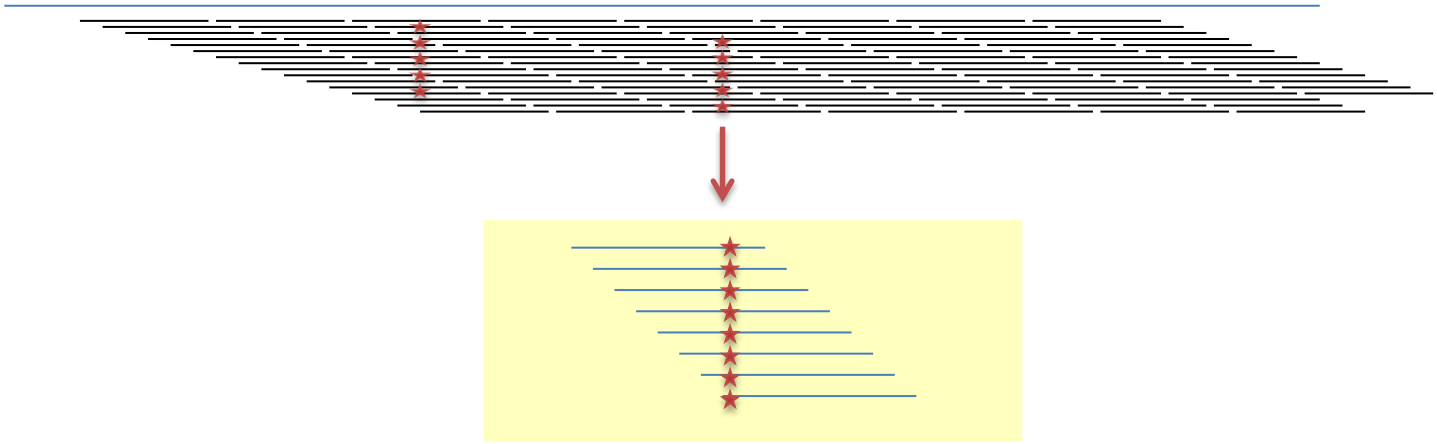
フィルター例



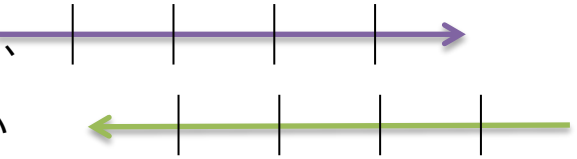
- Read direction filter 適用例：リードの色は緑（Forward）、赤（Reverse）、黄色（non-specific）を示しており、緑のリードが大部分のバリアントをサポートしていることがわかる。こういったアンバランスな箇所で検出されたバリアントが取り除かれる。

原理

- もしリードが理想的な均一なカバレッジであれば、検出されるバリアントをサポートする塩基のリード中の位置は、さまざまになるはずです。

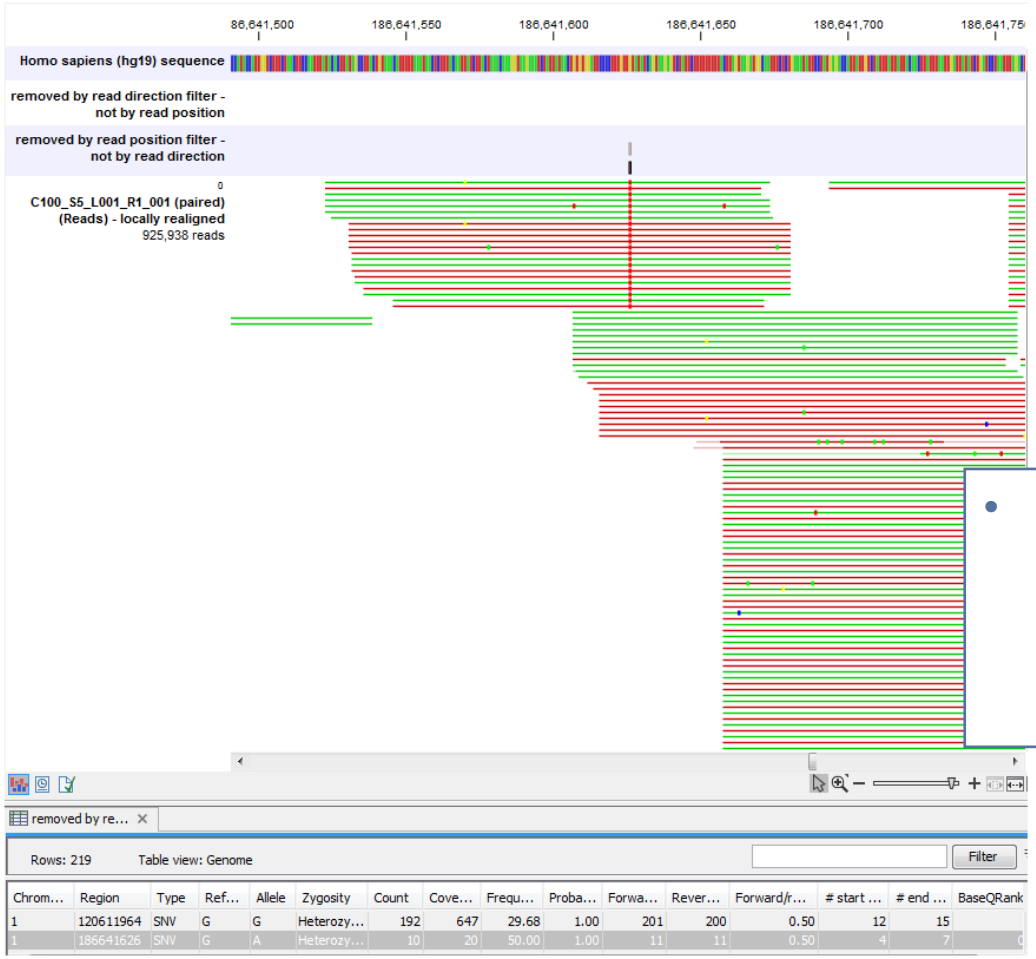


- これを使い、リードをForward、Reverseの向きを考慮して、それぞれ5分割、計10個の領域に分断し、変異が見つかった箇所がリードのどの領域に属するか、それらの分布が全体と大きく差がないかを検定しています



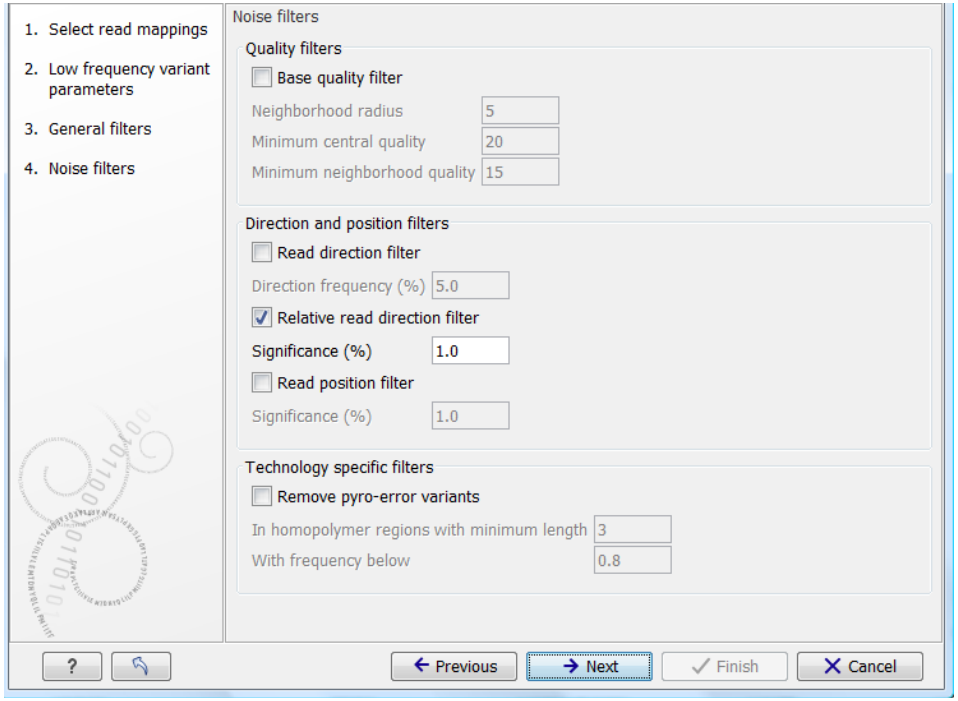
Read position filter

フィルター例

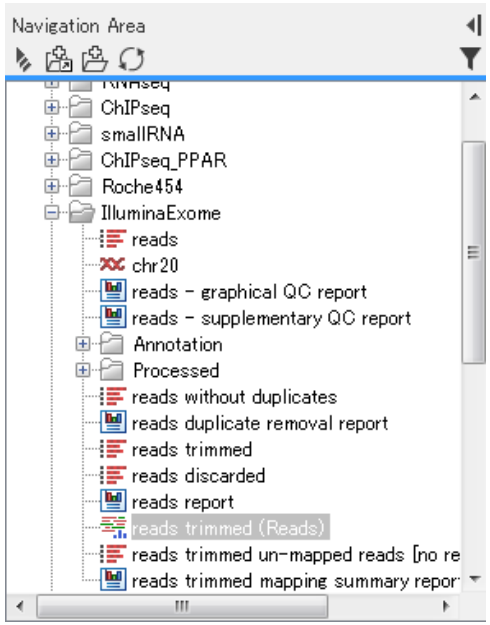


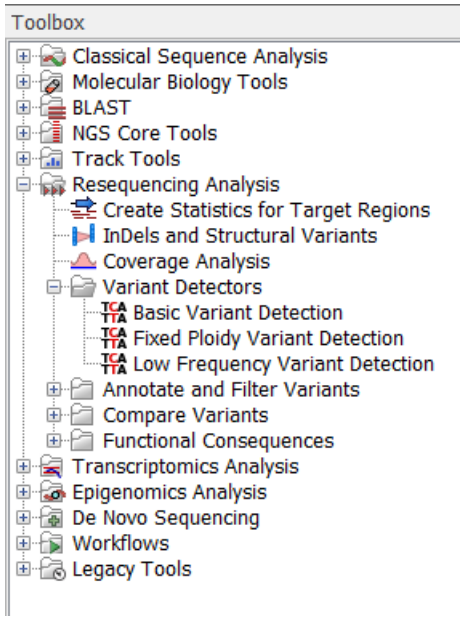
- Read position filter 適用例：バリアントをサポートしているリードがリードの同じ位置で検出されているため、このバリアントはRead position filter により除去されます。

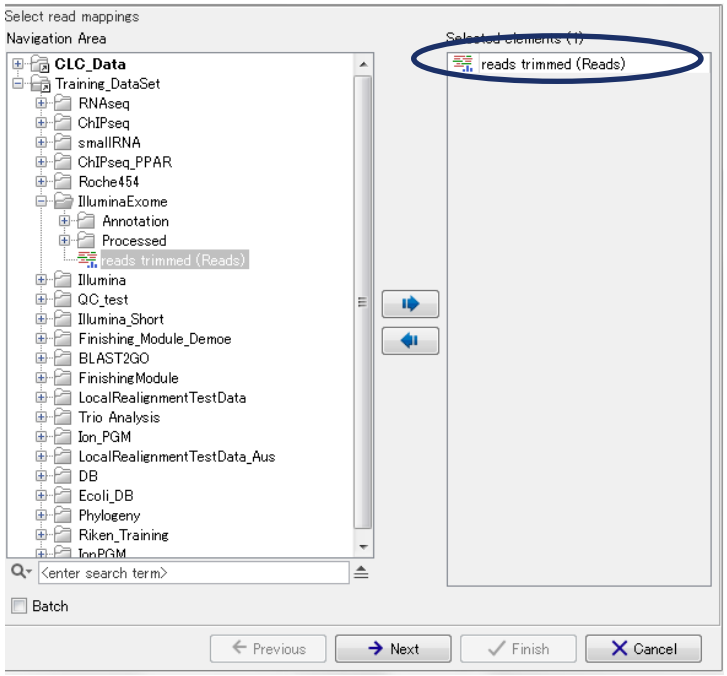
変異検出ツールでは、フィルターの条件をクリアした場合に variant をコールします



Basic Variant Detection

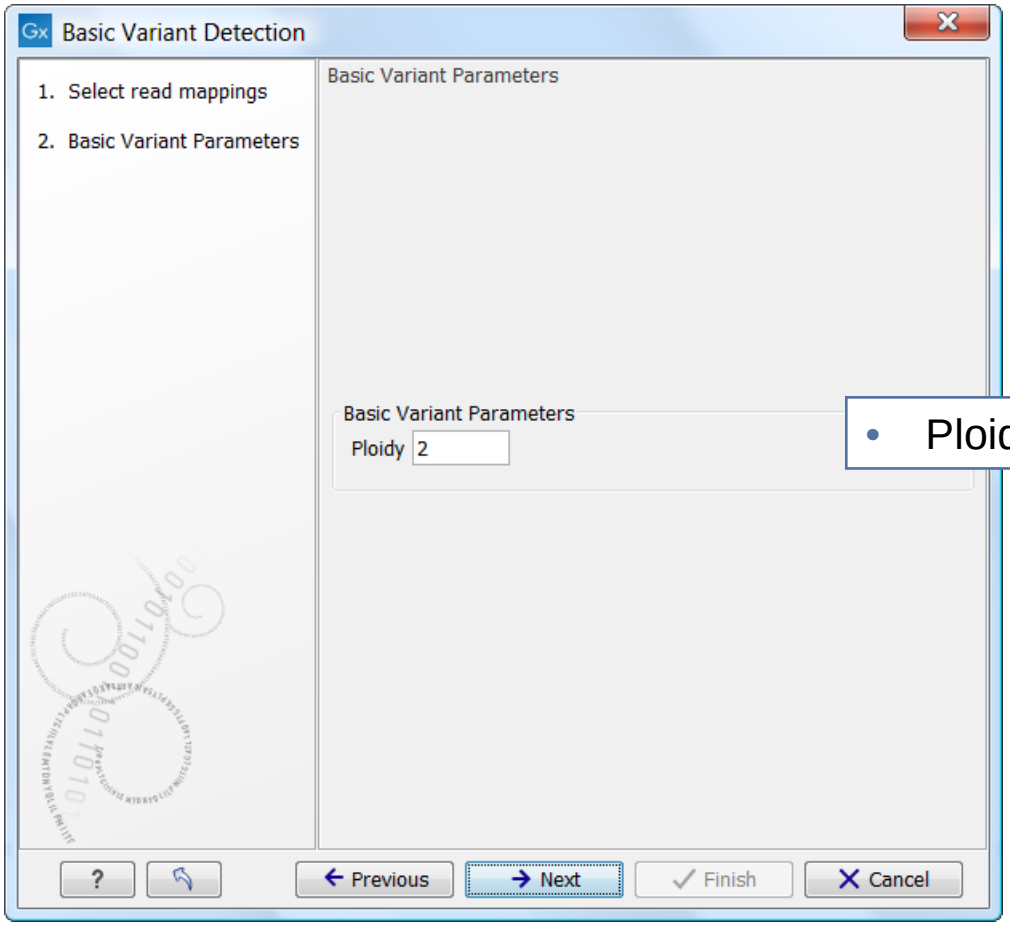




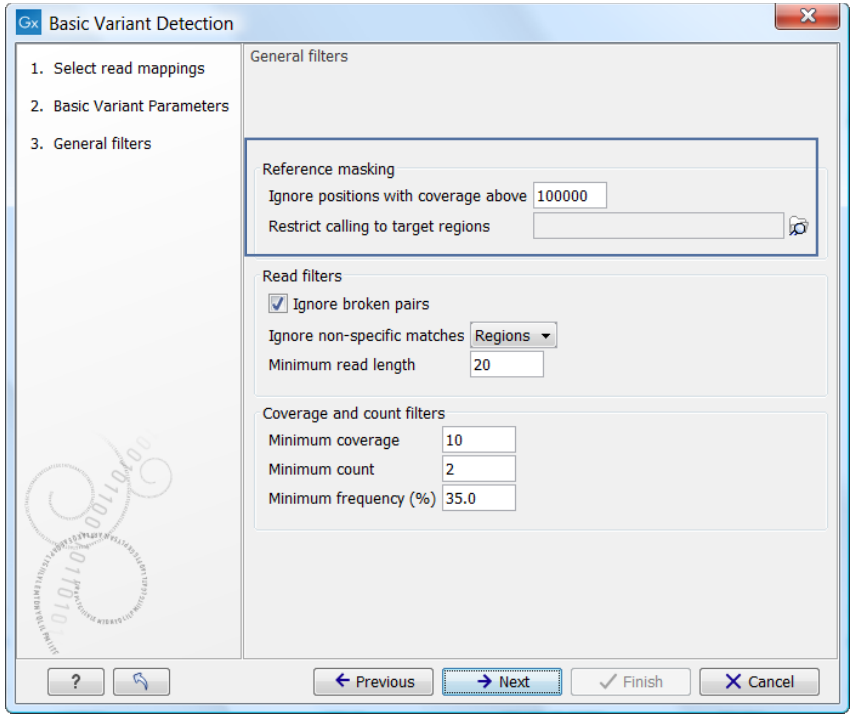


- Navigation Areaからマッピングデータを選択。
- Toolboxから Resequencing Analysis > Variant Detectors > Basic Variant Detection を選択、ダブルクリック。
- ウィザードが起動し、選択したデータが選ばれていることを確認。

Basic Variant Detection

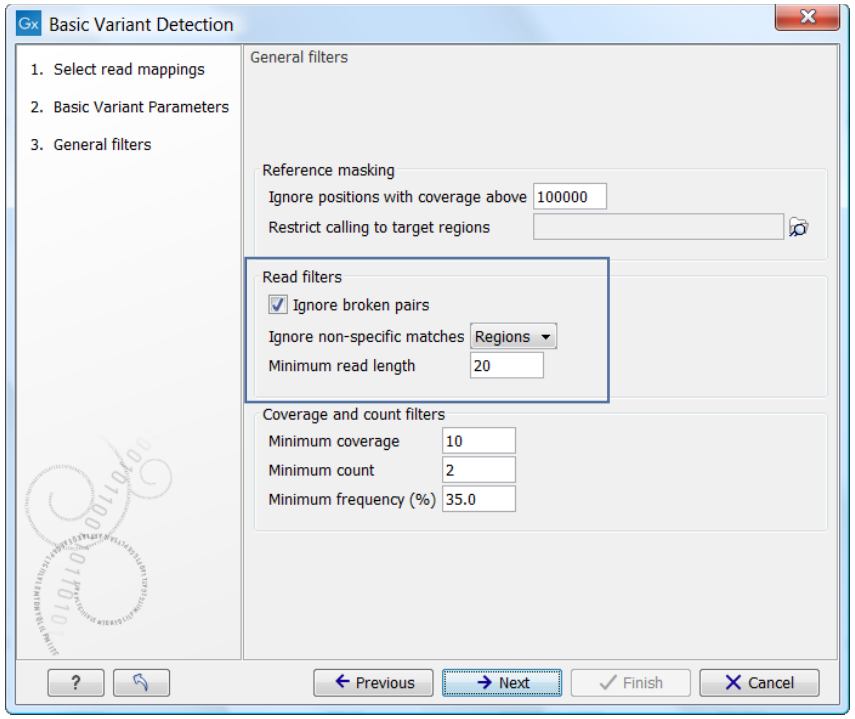


- Ploidy: 参照配列の倍数性



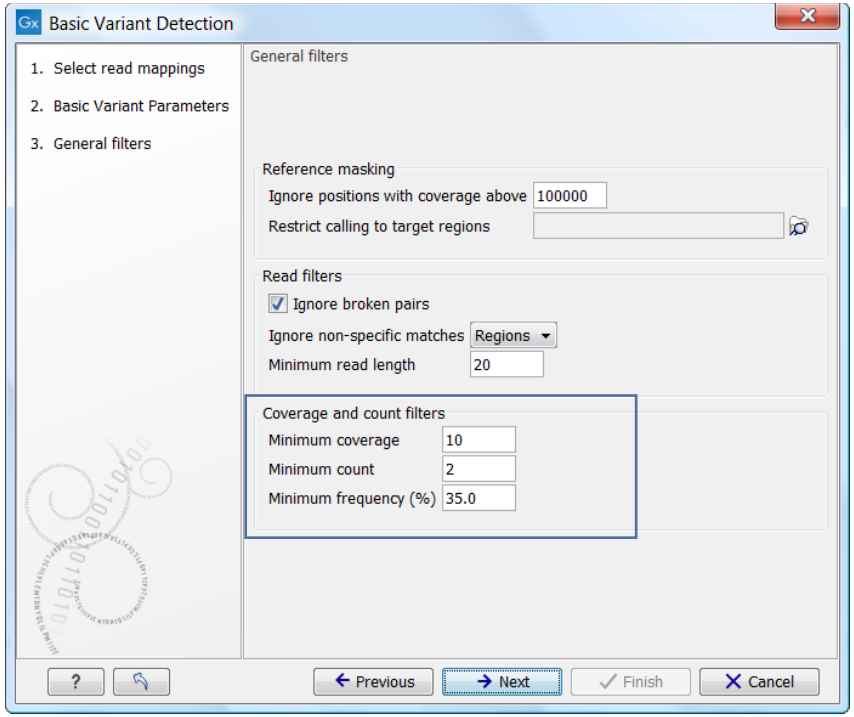
Reference masking

- Ignore positions with coverage above：カバレッジが指定した数字以上のバリアントについてリストに含めない
- Restrict calling to target regions：バリアントを検出したい領域の指定（アノテーショントラックで指定）



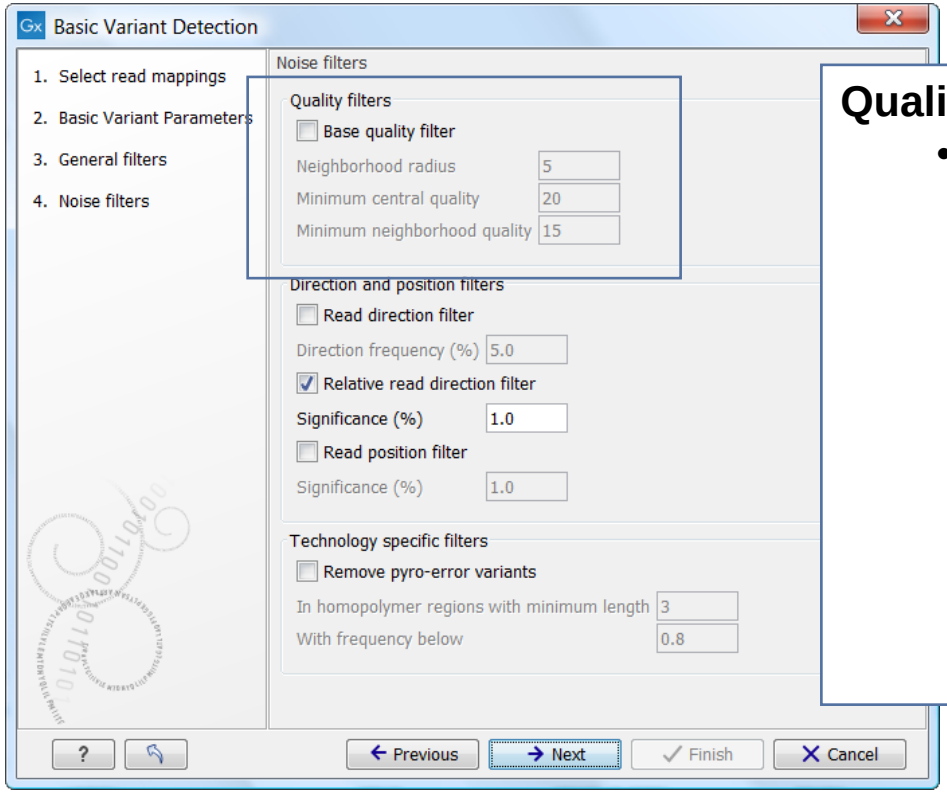
Read filters

- Ignore broken pairs：ペアエンドのリードでペアと認識されなかったリードをバリアント検出の計算に含めるかどうか
- Ignore non-specific matches：「Reads」を選択すると、non-specificなマッチのリードを計算に含めなくなり、「Regions」を選択すると、1本でもnon-specificなリードが含まれる場合、その領域のバリアントを検出しません。
- Minimum read length：Ignore broken pairとIgnore non-specific regions が指定された場合、このフィルターの対象となる最小のリードの長さの設定が必要です。これは非常に短いリードは、その短さからnon-specificになる可能性があるためです。



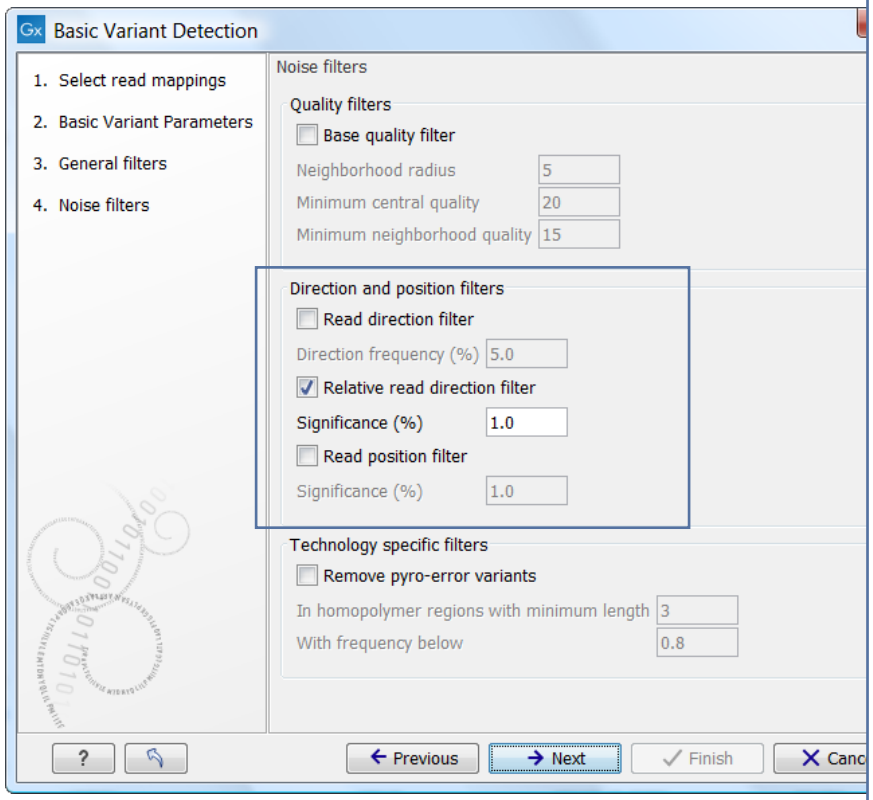
Coverage and count filters

Minimum coverage：最小カバレッジ
Minimum count：バリアントを支持するリードの最低カウント数
Minimum frequency (%)：最小頻度



Quality filter

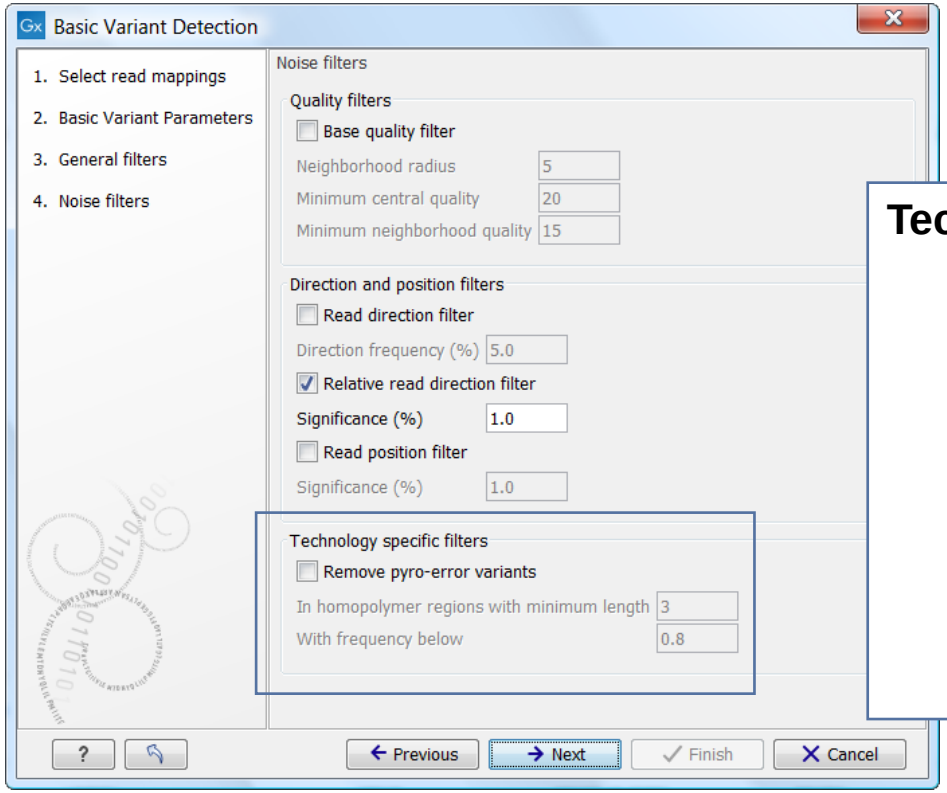
- Base quality filter:塩基のクオリティに関するフィルター
  - Neighborhood radius：クオリティフィルターの対象とする横方向の塩基数（奇数）
  - Minimum central quality：縦方向の数（リード数）
  - Minimum neighborhood quality：Neighborhood radiusで指定した範囲の最低クオリティ（Phred score）



Direction and position filters：

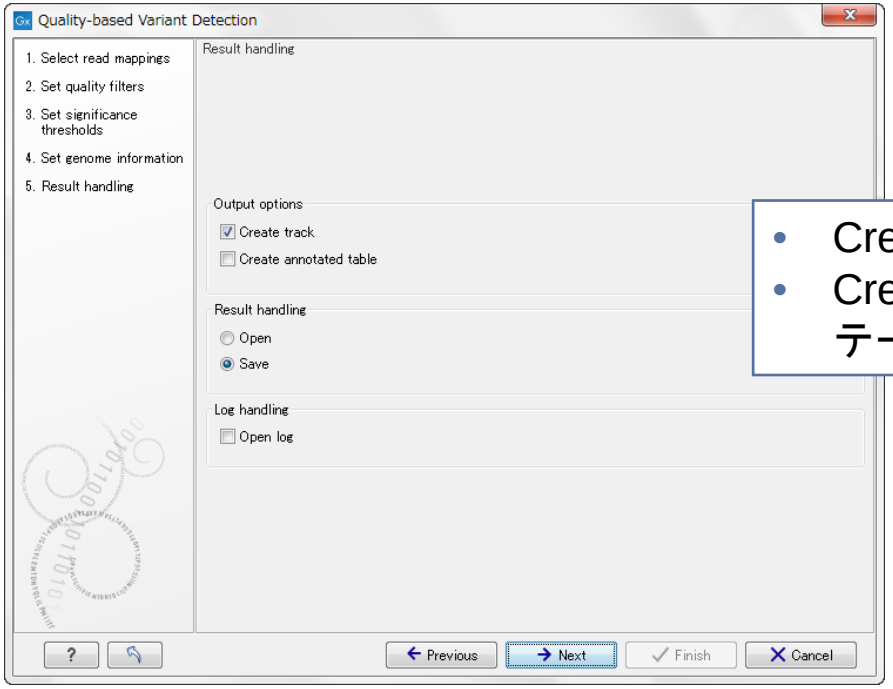
- Read direction filter：どちらか一方の方向のリードが多数見られる場合にそれを排除（ただし、アンプリコンには適していません）。
  - Relative read direction filter：リードの方向が一方のみに偏りすぎていないか、全体のForwardとReverseのバランスを見て統計検定を行う。Significanceで閾値を入力。
  - Read position filter：システマティックなエラーを取り除くために用いるツールでハイブリダイゼーションを行った場合のデータに有効。リードを5つのセグメントに分割し、バリアントの見られるポジションの５つのセグメントに分割されたリードの分布が全体のそれと似ているかどうか検定を行う。Significanceで閾値を入力。

Basic Variant Detection



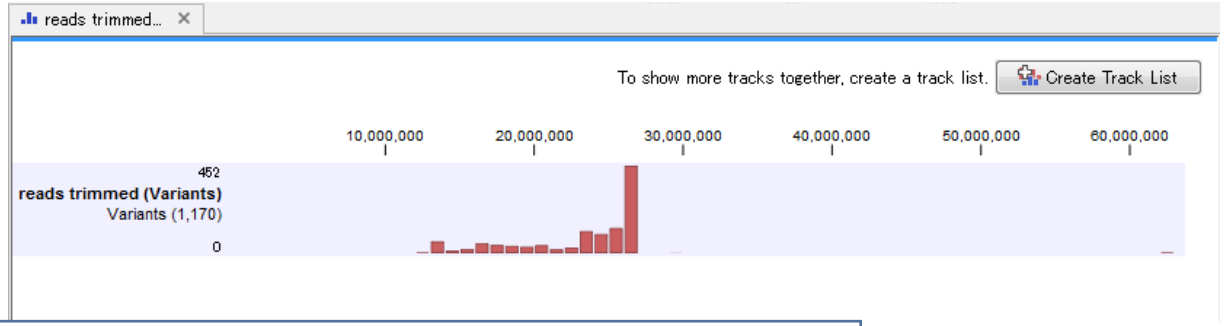
Technology specific filters

- Remove pyro-error variants：ホモポリマー領域に対するエラーの除去
  - In homopolymer regions with minimum length：指定した長さのホモポリマー領域のInDelを取り除く。
  - With frequency below：指定した頻度以下のものについてのみフィルターを適用。

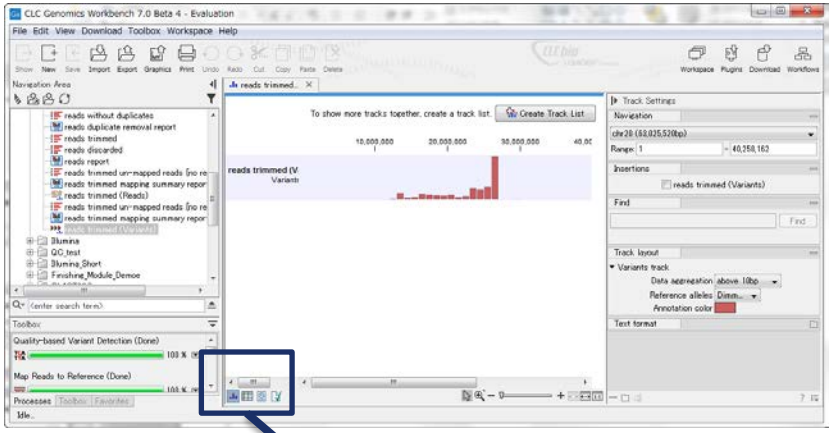


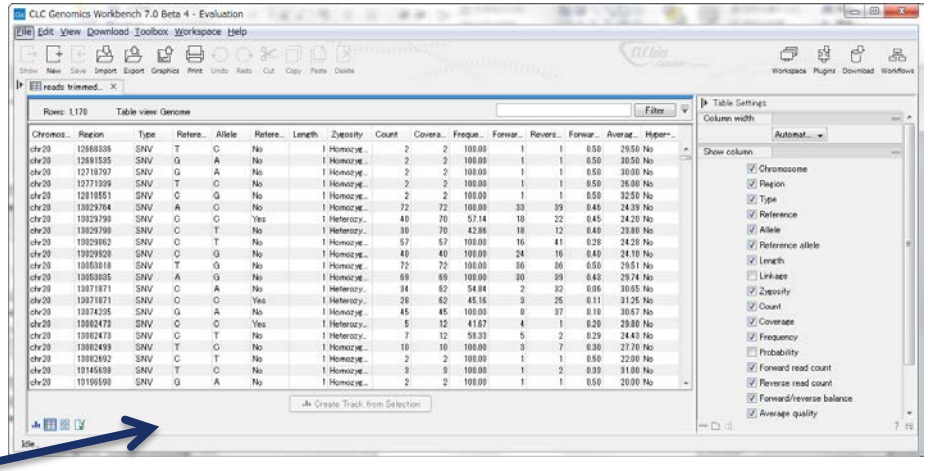
- Create track: トラックの作成
- Create annotated table: アノテーション付のテーブルの作成

結果



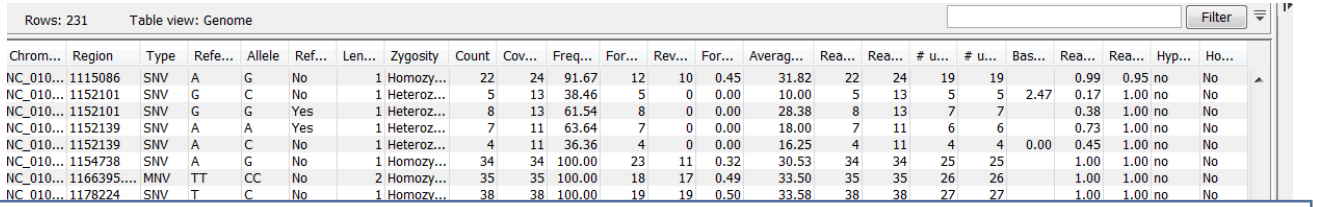
- 結果はデフォルトではトラックフォーマットになっています。





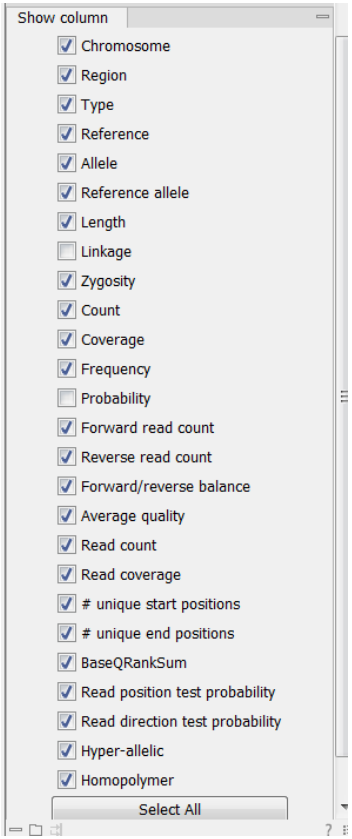
- 左下のテーブルアイコンをクリックするとテーブルに代わります。

Basic Variant Detection：結果



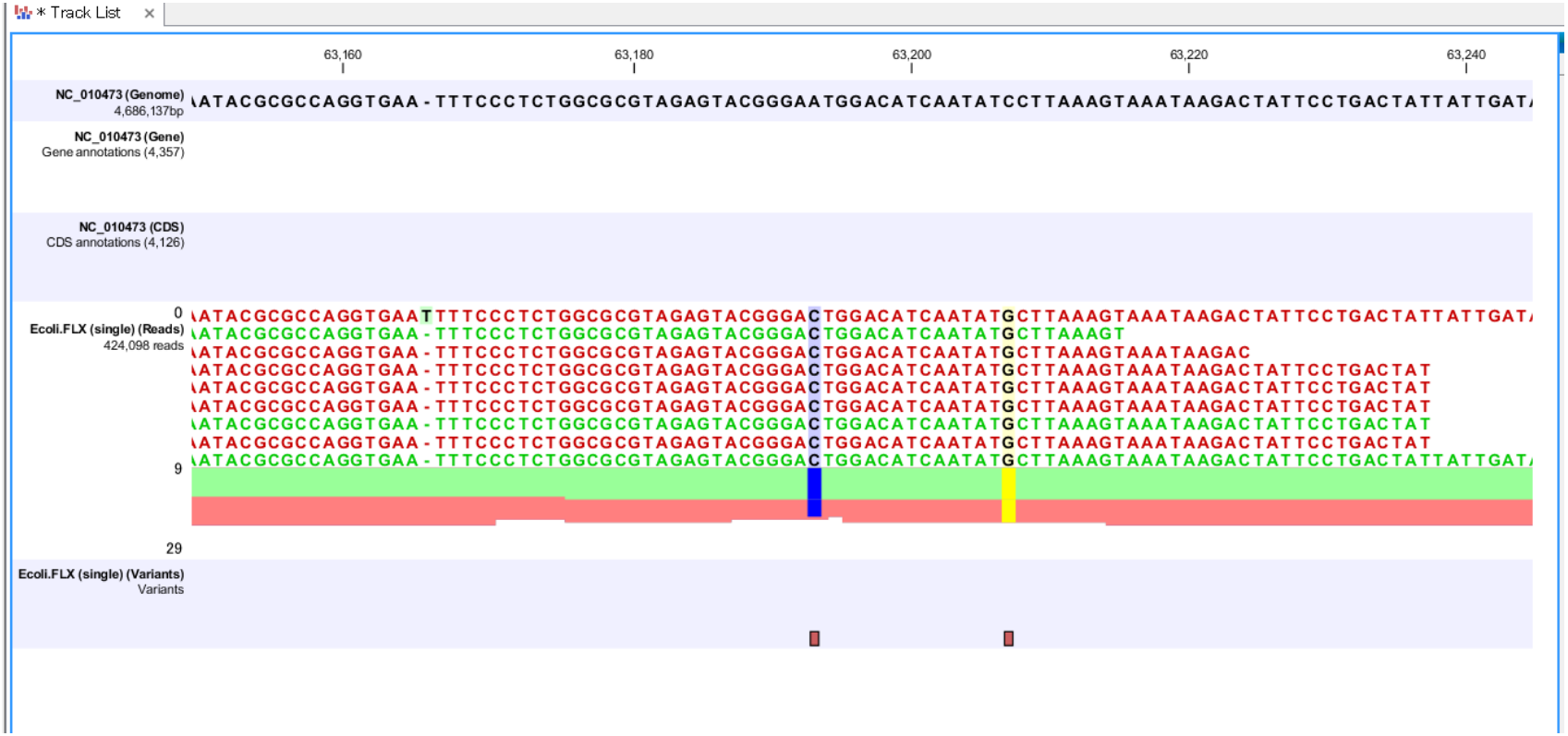
Rows: 231　Table view: Genome　Filter

Chrom...	Region	Type	Refe...	Allele	Ref...	Len...	Zygosity	Count	Cov...	Freq...	For...	Rev...	For...	Averag...	Rea...	Rea...	# u...	# u...	Bas...	Rea...	Rea...	Hyp...	Ho...
NC_010...	1115086	SNV	A	G	No	1	Homozy...	22	24	91.67	12	10	0.45	31.82	22	24	19	19		0.99	0.95	no	No
NC_010...	1152101	SNV	G	C	No	1	Heteroz...	5	13	38.46	5	0	0.00	10.00	5	13	5	5	2.47	0.17	1.00	no	No
NC_010...	1152101	SNV	G	G	Yes	1	Heteroz...	8	13	61.54	8	0	0.00	28.38	8	13	7	7		0.38	1.00	no	No
NC_010...	1152139	SNV	A	A	Yes	1	Heteroz...	7	11	63.64	7	0	0.00	18.00	7	11	6	6		0.73	1.00	no	No
NC_010...	1152139	SNV	A	C	No	1	Heteroz...	4	11	36.36	4	0	0.00	16.25	4	11	4	4	0.00	0.45	1.00	no	No
NC_010...	1154738	SNV	A	G	No	1	Homozy...	34	34	100.00	23	11	0.32	30.53	34	34	25	25		1.00	1.00	no	No
NC_010...	1166395....	MNV	TT	CC	No	2	Homozy...	35	35	100.00	18	17	0.49	33.50	35	35	26	26		1.00	1.00	no	No
NC_010...	1178224	SNV	T	C	No	1	Homozy...	38	38	100.00	19	19	0.50	33.58	38	38	27	27		1.00	1.00	no	No

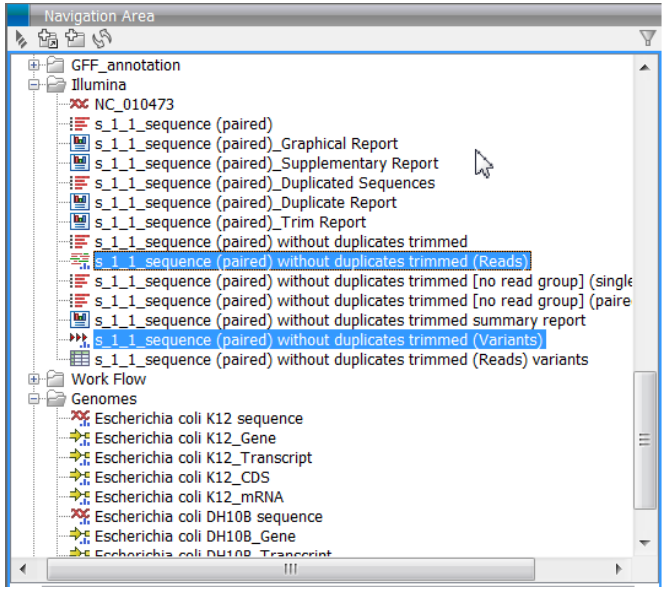


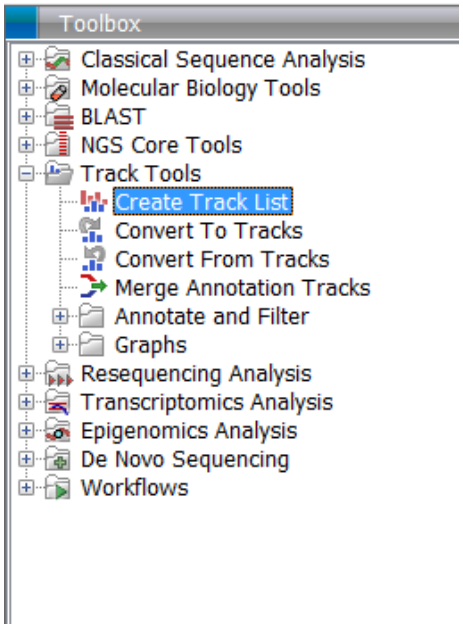
- Count: クオリティのフィルターをパスしたリードの数
- Coverage: クオリティのフィルターをパスしたリードの数
- Frequency: バリアントが見られた頻度
- Probability: バリアントのアレルの事後確率（そのアレルが尤もであるとする確率。高い方がより確度が高いという事。）
- Forward reads: その領域に見られたForwardリードの数
- Reverse reads:その領域に見られたReverseリードの数
- Forward/reverse: Forward/Total reads　または　Reverse/Total reads のうち小さい方の値。ForwardとReverseが同じなら、0.5となる。
- Average quality: 該当する領域の平均リードクオリティ。
- # unique start positions:バリアントコールに使われたリードのうちスタートポジションにあるリードの数
- # unique end positions:バリアントコールに使われたリードのうち最後の箇所にあるリードの数
- BaseQRankSum:クオリティスコアについて、参照配列と同じアレルとバアリアントのアレルについてマンホイットニーU検定を行い計算されたZスコア。これが高いほど参照配列の塩基とバリアントの塩基に差がある。
- Hyper-alleic：想定されるアレルよりも頻度が高いかどうか
- Homopolymer：ホモポリマー領域かどうか

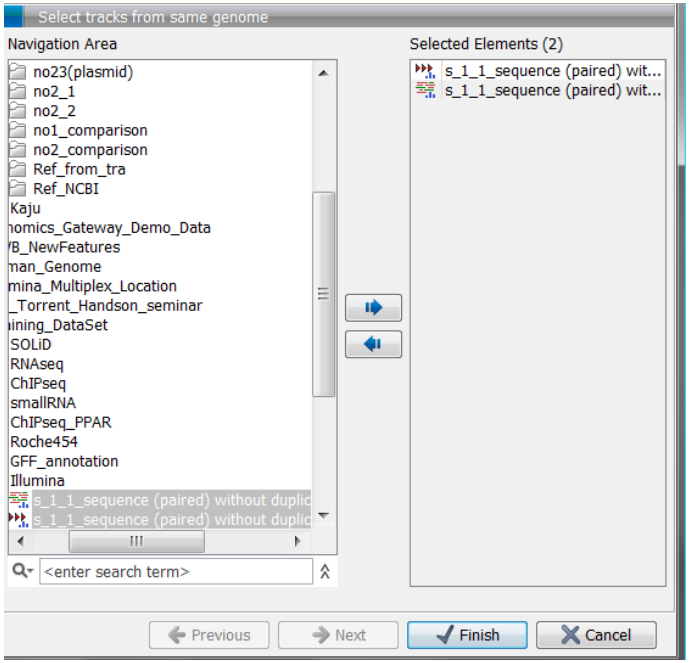
トラックリストの作成


* Track List
63,160
63,180
63,200
63,220
63,240
NC_010473 (Genome)
4,686,137bp
NC_010473 (Gene)
Gene annotations (4,357)
NC_010473 (CDS)
CDS annotations (4,126)
0
Ecoli.FLX (single) (Reads)
424,098 reads
9
29
Ecoli.FLX (single) (Variants)
Variants

トラックリスト作成

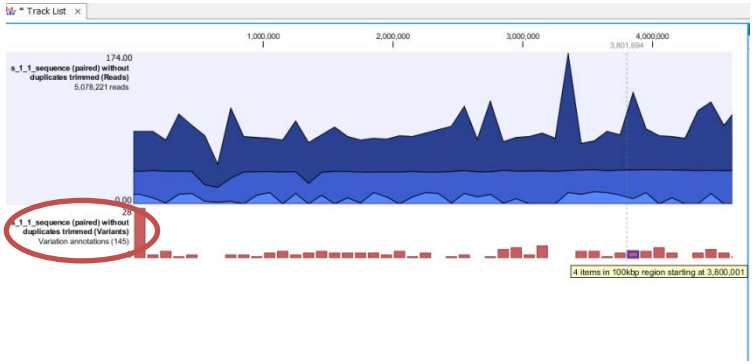




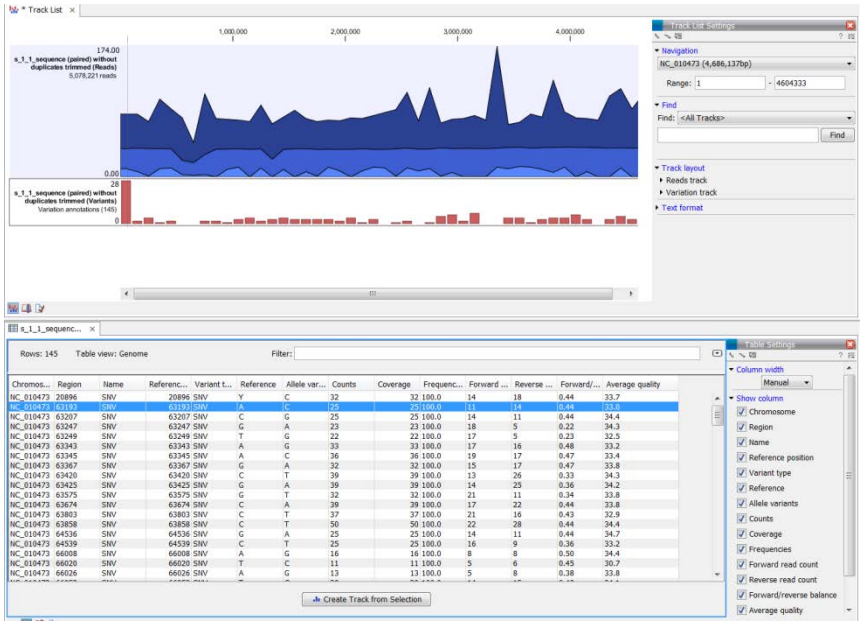


- Navigation Areaからマッピングデータとバリアントの結果を選択。
- Toolboxから ResequTrack Tools > Create Track List を選択、ダブルクリック。
- ウィザードが起動し、選択したデータが選ばれていることを確認。

トラックリストの作成



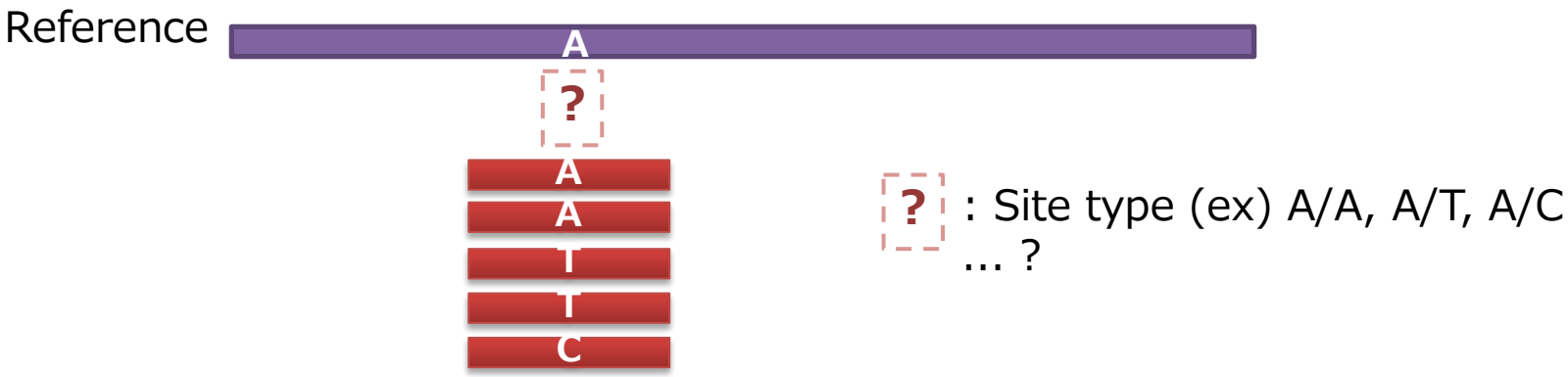
- バリアントのトラックの名前のところでダブルクリック



- テーブルが現れます。テーブルの行と、マッピングのビューアは対応しているので、テーブルで指定したポジションに自動的にビューアが移動します。

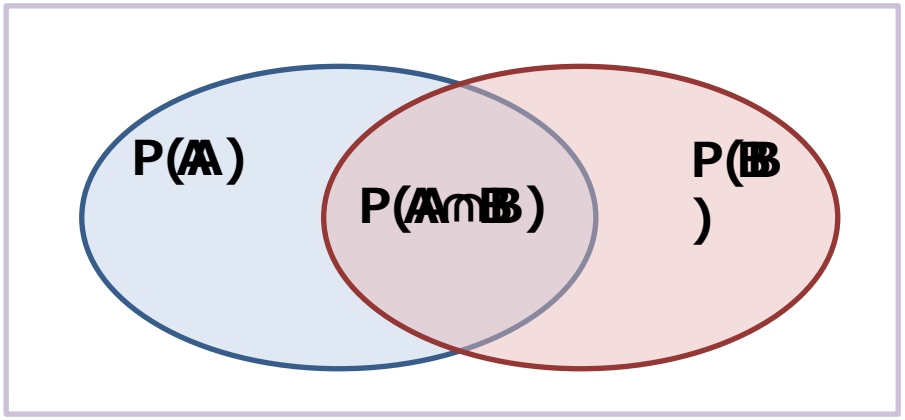
Probabilistic Variant Detection

- 確率モデル（Bayes model）を使ったバリアント検出



与えられるリードから、そのポジションのSite Typeを推定
Reference と推定したSite typeが異なる場合、バリアントとして結果返す。

Fixed Ploidy Variant Detection



$$P(A \cap B) = P(B \mid A)P(A)$$
$$P(A \cap B) = P(A \mid B)P(B)$$
$$P(B \mid A)P(A) = P(A \mid B)P(B)$$

ベイズの定理

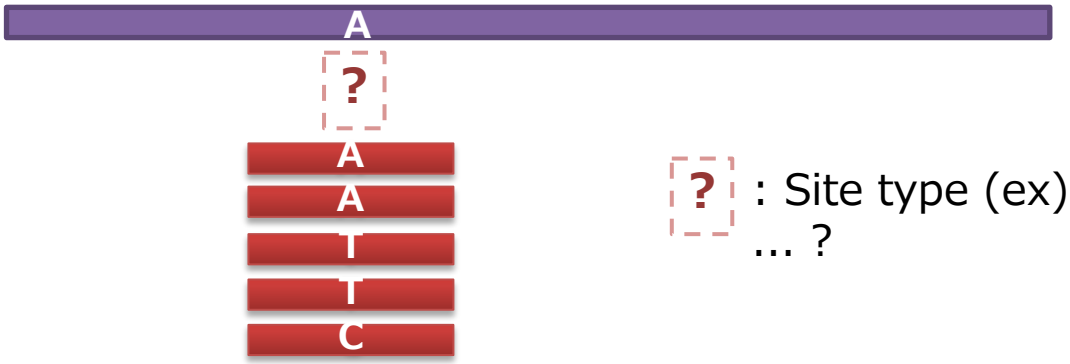
$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$$

事後確率
Posterior

事前確率
Prior

尤度
Likelihood

Reference



? : Site type (ex) A/A, A/T, A/C … ?

$$P(S \mid R) = \frac{P(R \mid S)P(S)}{P(R)}$$

S : Site type

R : Reads

$P(R \mid S)$: **Error Model を使って推定**

$P(S)$: **Genome Model を使って推定**

Genome Model

Reference がAのとき、Readの大部分はAになると仮定し、初期の確率を以下のように設定し、EMアルゴリズムを使ってそれぞれの確率を推定する。

EMアルゴリズム（Expectation Maximization algorithm）は、得られたデータから推定したい現象が観察できない場合に、その確率を推定する、一般的な統計の手法。

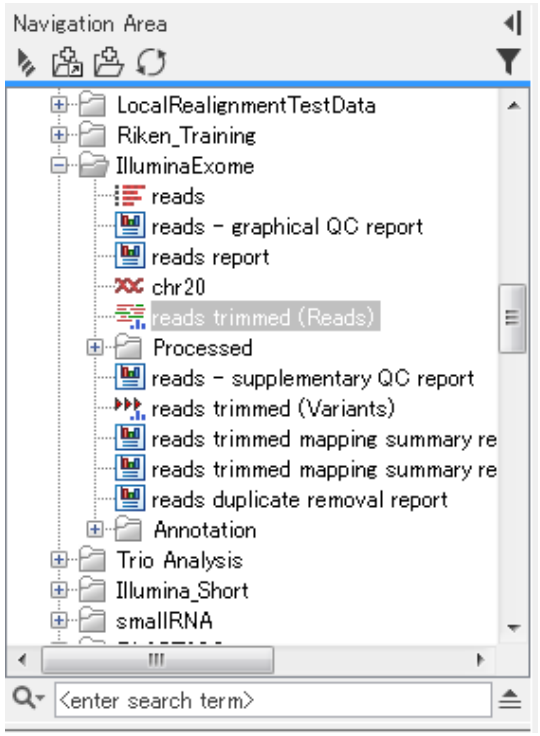
Site Type	Initial Probability
A/A	0.2475
A/C	0.001
A/G	0.001
A/T	0.001
T/C	0.001
T/G	0.001
T/T	0.2475
G/C	0.001
C/C	0.2475
G/G	0.2475
G/-	0.001
A/-	0.001
C/-	0.001
T/-	0.001

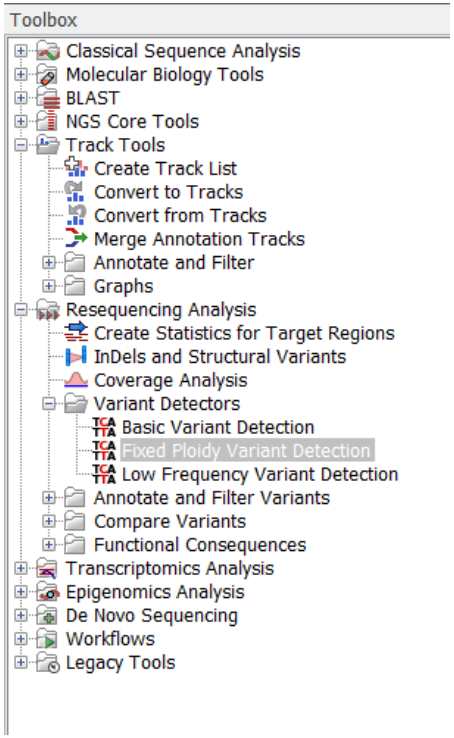
Error Model

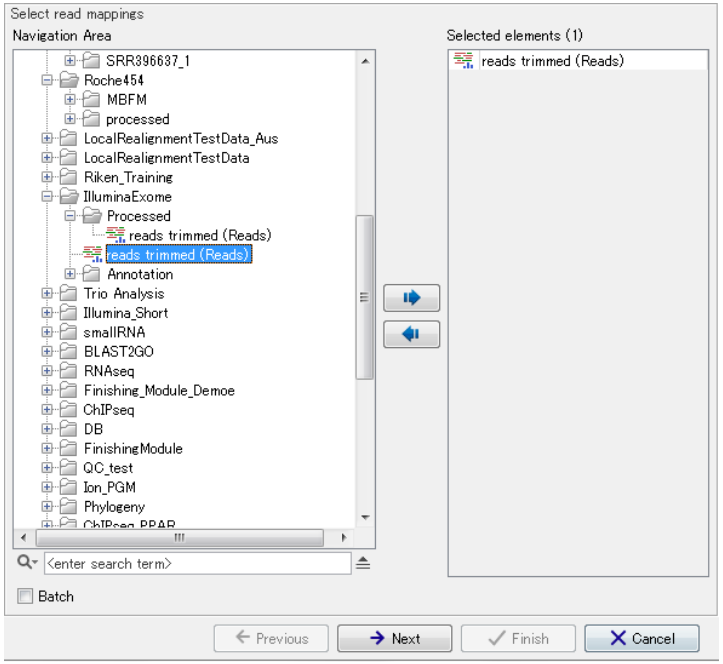
– リードに含まれるエラーを考慮するため、尤度のところにエラーを考慮した確率を推定する。初期値を以下のように設定し、EMアルゴリズムにて確率を推定する。

Reads \ Reference	A	C	G	T	-
A	0.90	0.025	0.025	0.025	0.025
C	0.025	0.90	0.025	0.025	0.025
G	0.025	0.025	0.90	0.025	0.025
T	0.025	0.025	0.025	0.90	0.025
-	0.025	0.025	0.025	0.025	0.90

Fixed Ploidy Variant Detection

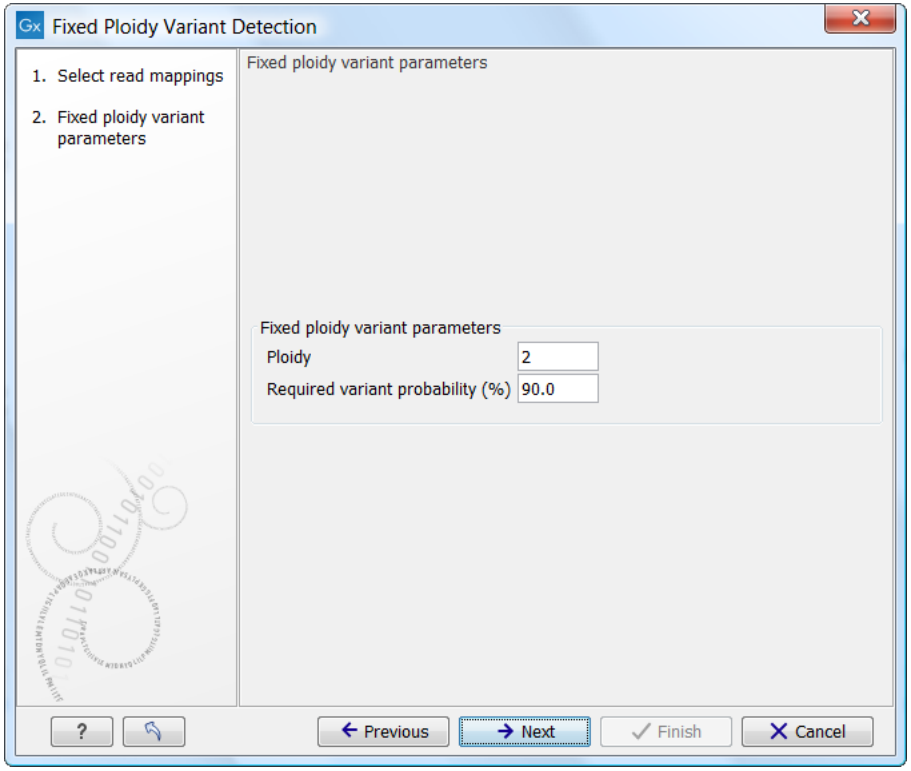






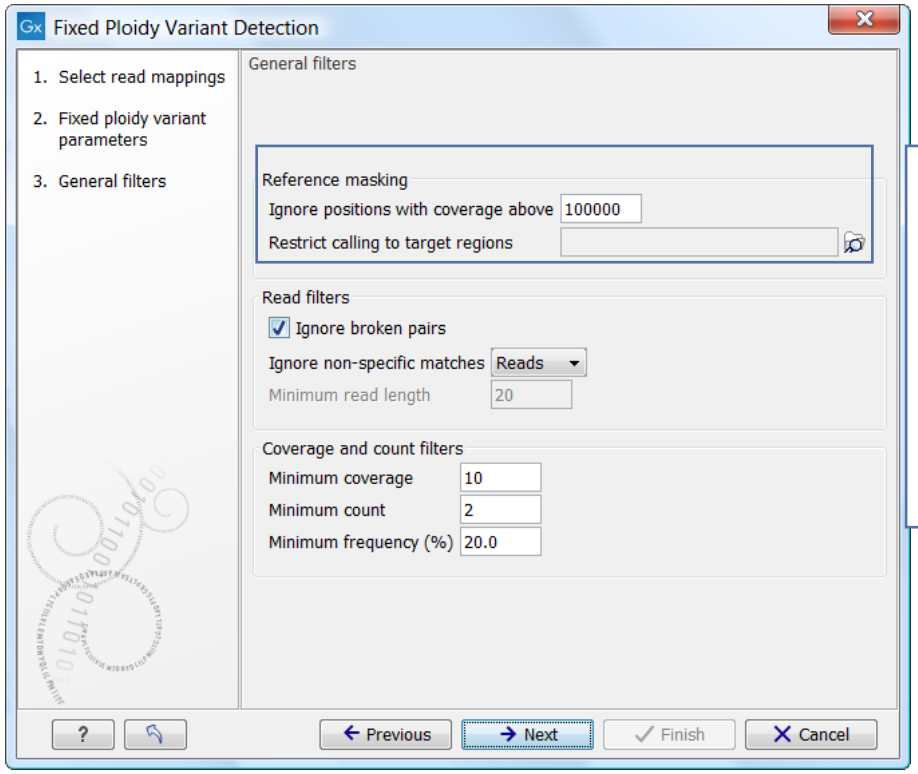
- Navigation Areaからマッピングデータを選択。
- Toolboxから Resequencing Analysis > Variant Detectors > Fixed Ploidy Variant Detection を選択、ダブルクリック。
- ウィザードが起動し、選択したデータが選ばれていることを確認。

Fixed Ploidy Variant Detection



Ploidy：参照配列の倍数性

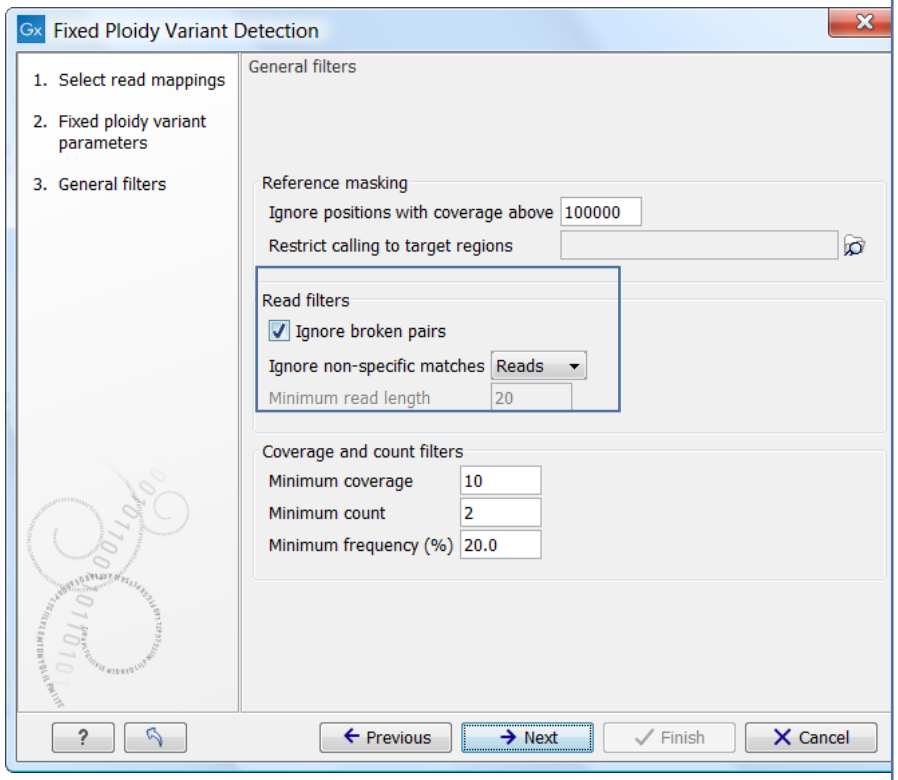
- Required variant probability：バリアントが参照配列と異なる確率（想定で入力）。この値を低くすると、検出されるバリアントが多くなります。



Reference masking

- Ignore positions with coverage above：カバレッジが指定した数字以上のバリアントについてリストに含めない
- Restrict calling to target regions：バリアントを検出したい領域の指定（アノテーショントラックで指定）

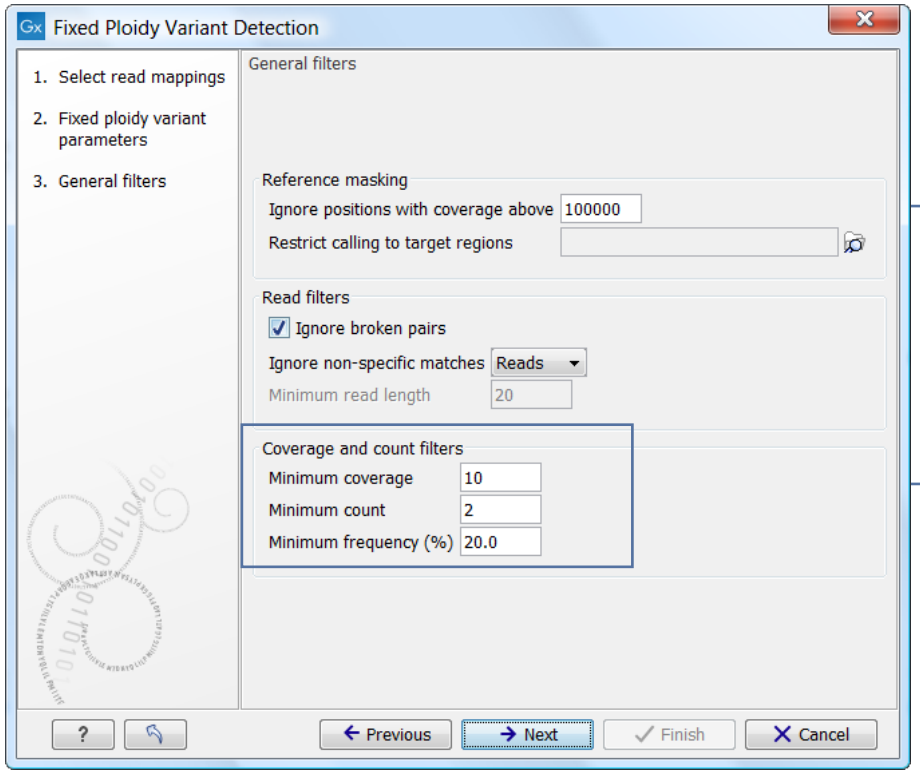
Fixed Ploidy Variant Detection



Read filters

- Ignore broken pairs：ペアエンドのリードでペアと認識されなかったリードをバリアント検出の計算に含めるかどうか
- Ignore non-specific matches：「Reads」を選択すると、non-specificなマッチのリードを計算に含めなくなり、「Regions」を選択すると、1本でもnon-specificなリードが含まれる場合、その領域のバリアントを検出しません。
- Minimum read length：Ignore broken pairとIgnore non-specific regions が指定された場合、このフィルターの対象となる最小のリードの長さの設定が必要です。これは非常に短いリードは、その短さからnon-specificになる可能性があるためです。

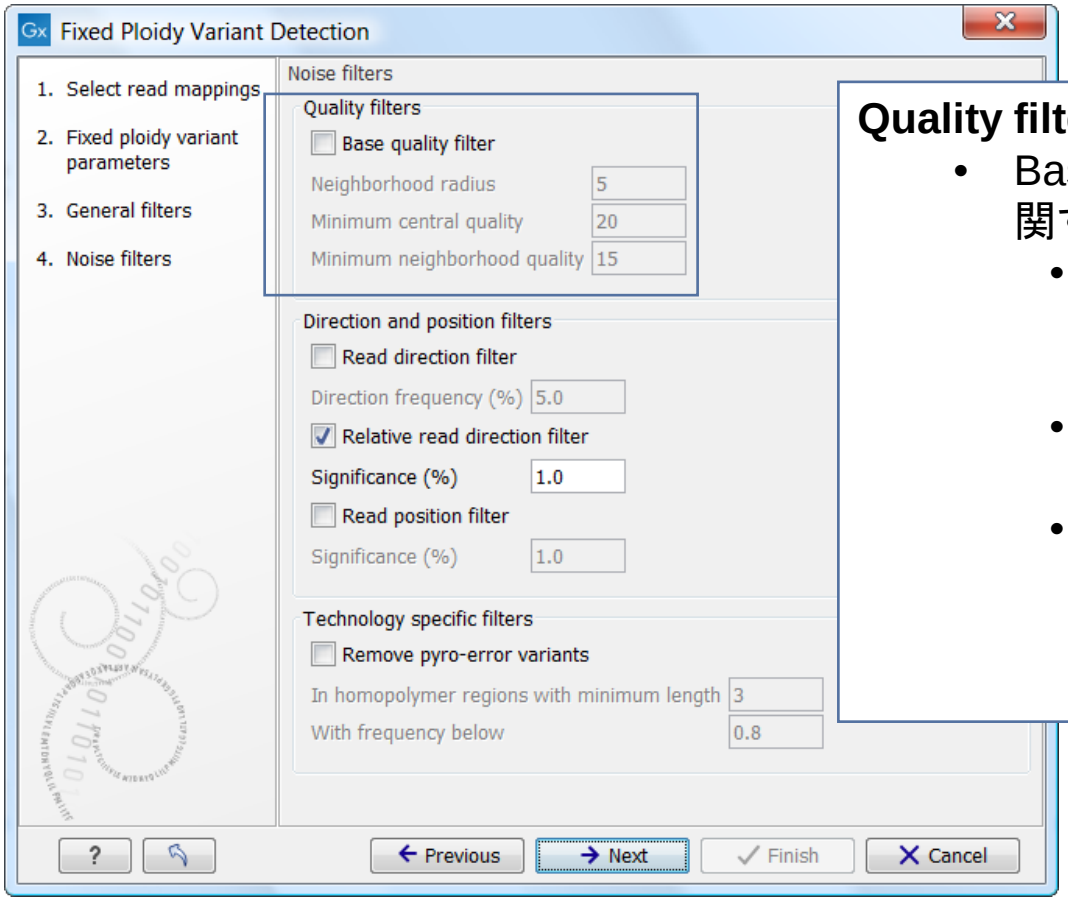
Fixed Ploidy Variant Detection



Coverage and count filters

Minimum coverage：最小カバレッジ
Minimum count：バリアントを支持するリードの最低カウント数
Minimum frequency (%)：最小頻度

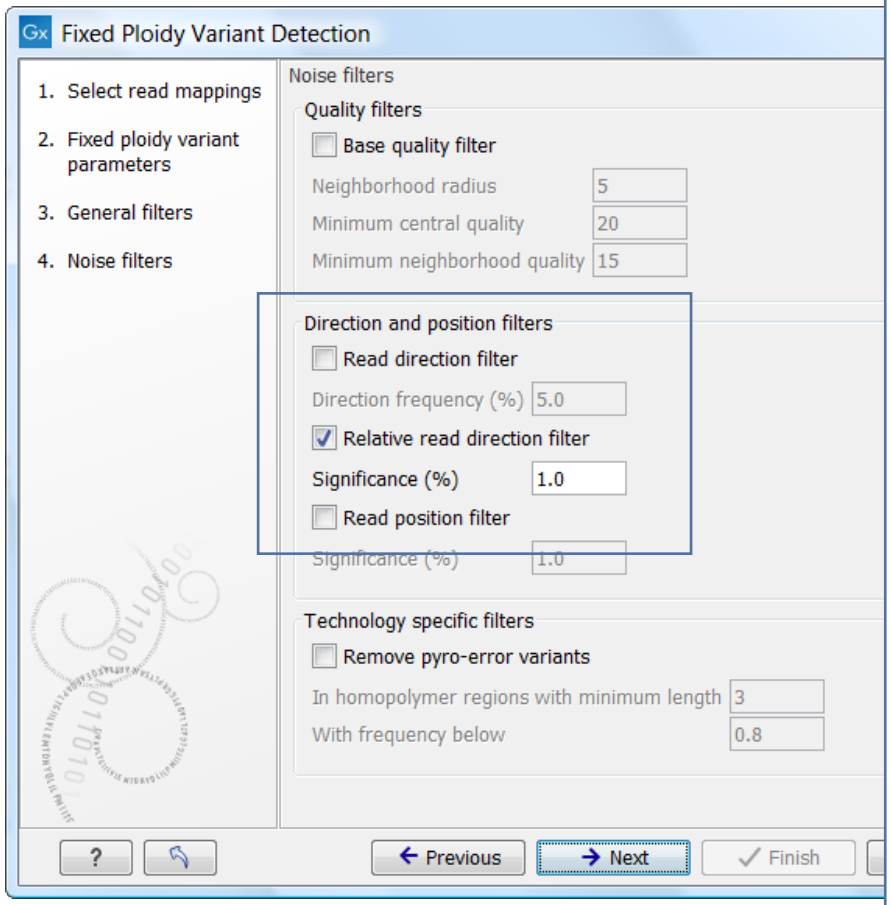
Fixed Ploidy Variant Detection



Quality filter

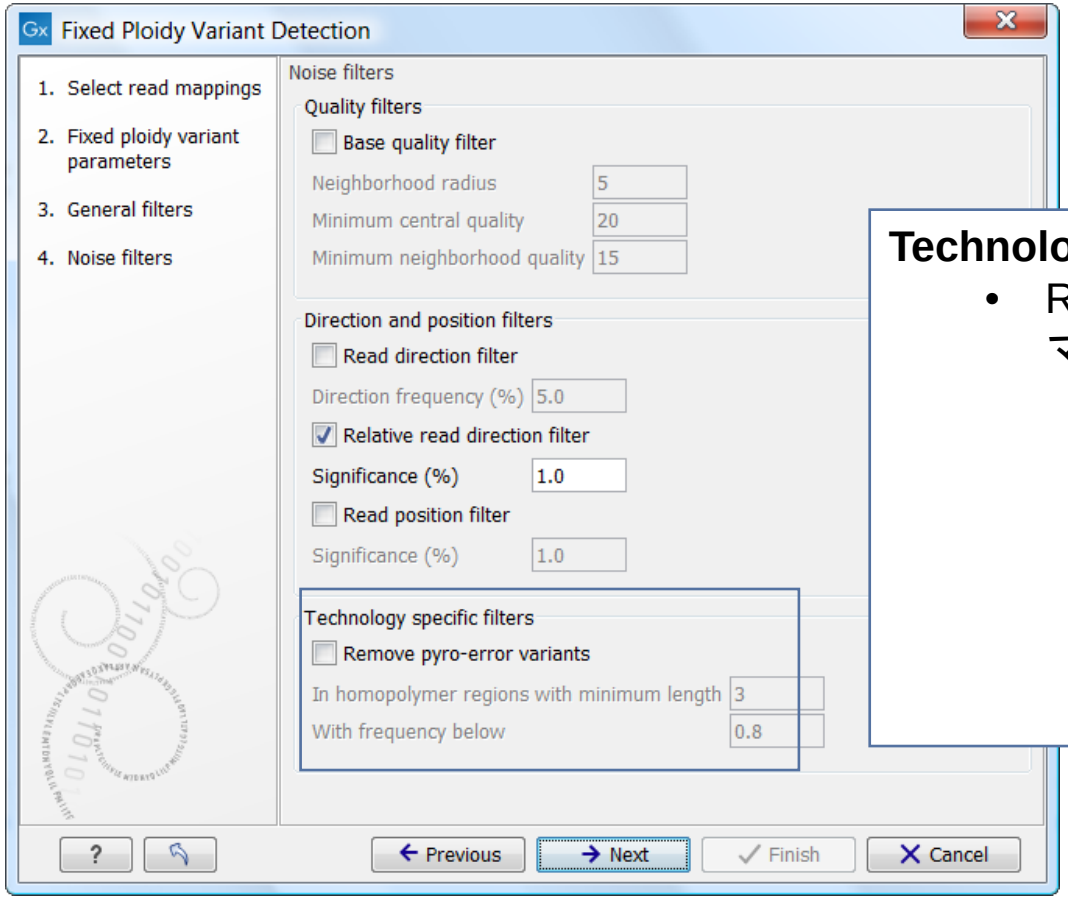
- Base quality filter:塩基のクオリティに関するフィルター
  - Neighborhood radius：クオリティフィルターの対象とする横方向の塩基数（奇数）
  - Minimum central quality：縦方向の数（リード数）
  - Minimum neighborhood quality：Neighborhood radiusで指定した範囲の最低クオリティ（Phred score）

Fixed Ploidy Variant Detection



Direction and position filters：

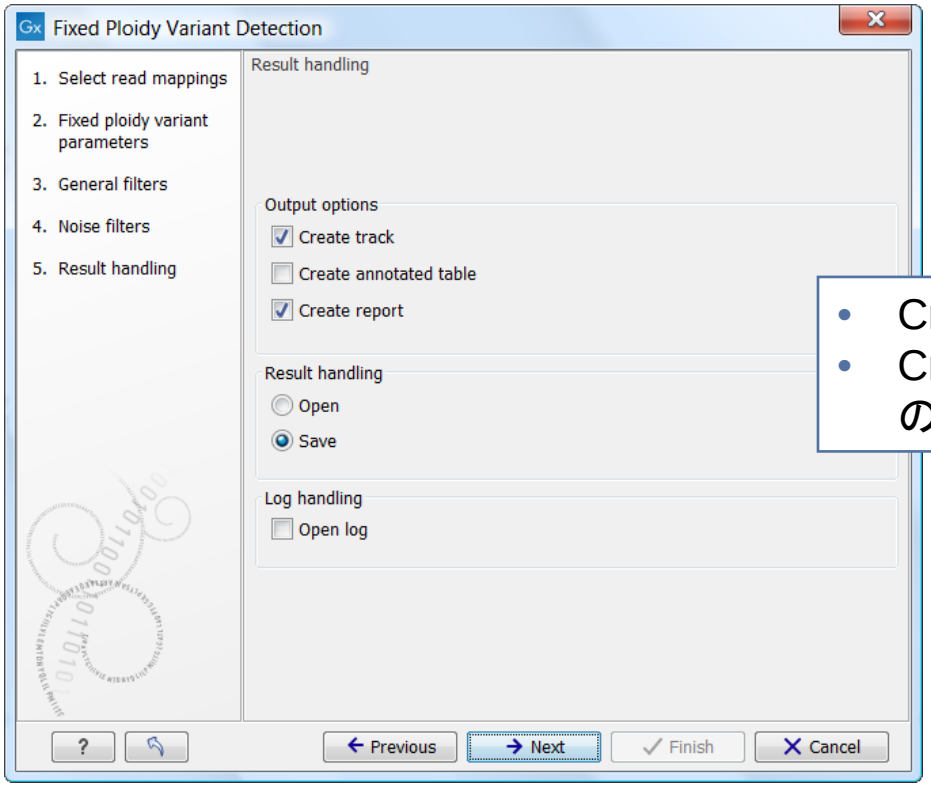
- Read direction filter：どちらか一方の方向のリードが多数見られる場合にそれを排除（ただし、アンプリコンには適していません）。
  - Relative read direction filter：リードの方向が一方のみに偏りすぎていないか、全体のForwardとReverseのバランスを見て統計検定を行う。Significanceで閾値を入力。
  - Read position filter：システマティックなエラーを取り除くために用いるツールでハイブリダイゼーションを行った場合のデータに有効。リードを5つのセグメントに分割し、バリアントの見られるポジションの５つのセグメントに分割されたリードの分布が全体のそれと似ているかどうか検定を行う。Significanceで閾値を入力。



Technology specific filters

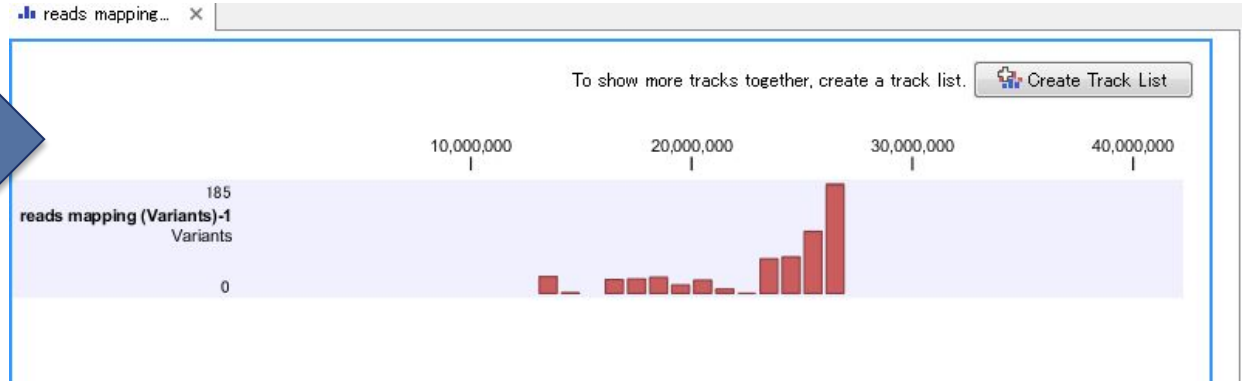
- Remove pyro-error variants：ホモポリマー領域に対するエラーの除去
  - In homopolymer regions with minimum length：指定した長さのホモポリマー領域のInDelを取り除く。
  - With frequency below：指定した頻度以下のものについてのみフィルターを適用。

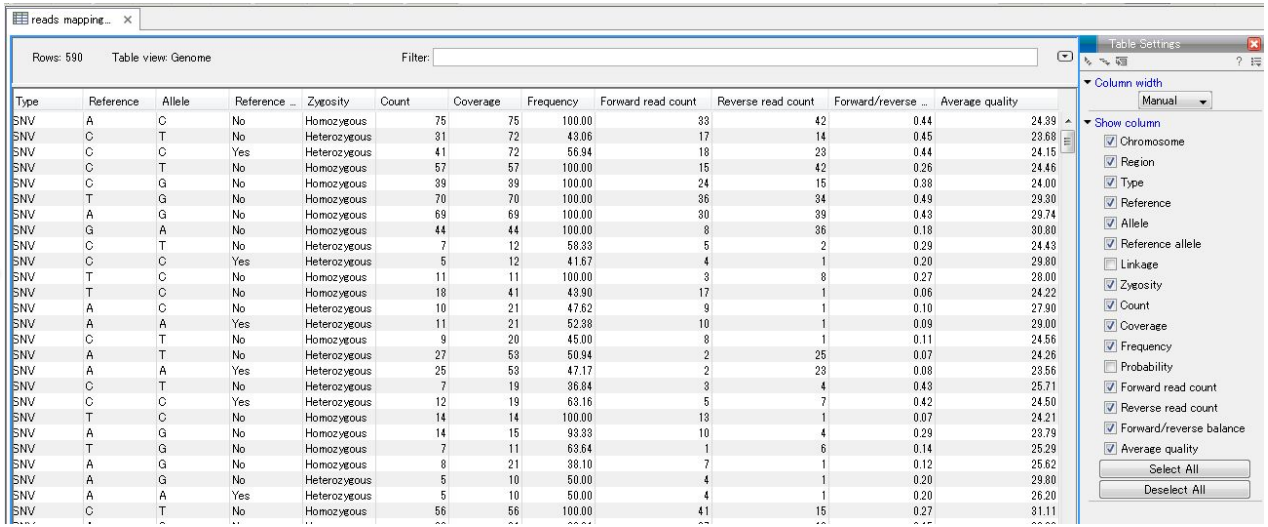
Fixed Ploidy Variant Detection



- Create track: トラックの作成
- Create annotated table: アノテーション付のテーブルの作成

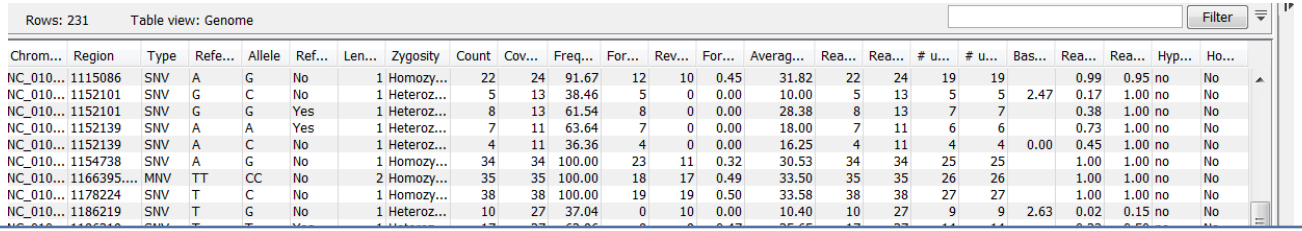
Fixed Ploidy Variant Detection：ビューの見方



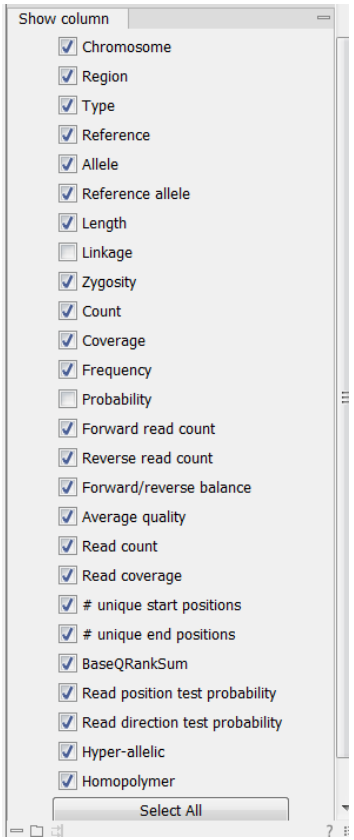


Type	Reference	Allele	Reference ...	Zygosity	Count	Coverage	Frequency	Forward read count	Reverse read count	Forward/reverse ...	Average quality
SNV	A	C	No	Homozygous	75	75	100.00	33	42	0.44	24.39
SNV	C	T	No	Heterozygous	31	72	43.06	17	14	0.45	23.68
SNV	C	C	Yes	Heterozygous	41	72	56.94	18	23	0.44	24.15
SNV	C	T	No	Homozygous	57	57	100.00	15	42	0.26	24.46
SNV	C	G	No	Homozygous	39	39	100.00	24	15	0.38	24.00
SNV	T	G	No	Homozygous	70	70	100.00	36	34	0.49	29.30
SNV	A	G	No	Homozygous	69	69	100.00	30	39	0.43	29.74
SNV	G	A	No	Homozygous	44	44	100.00	8	36	0.18	30.80
SNV	C	T	No	Heterozygous	7	12	58.33	5	2	0.29	24.43
SNV	C	C	Yes	Heterozygous	5	12	41.67	4	1	0.20	29.80
SNV	T	C	No	Homozygous	11	11	100.00	3	8	0.27	28.00
SNV	T	C	No	Homozygous	18	41	43.90	17	1	0.06	24.22
SNV	A	C	No	Heterozygous	10	21	47.62	9	1	0.10	27.90
SNV	A	A	Yes	Heterozygous	11	21	52.38	10	1	0.09	29.00
SNV	C	T	No	Homozygous	9	20	45.00	8	1	0.11	24.56
SNV	A	T	No	Heterozygous	27	53	50.94	2	25	0.07	24.26
SNV	A	A	Yes	Heterozygous	25	53	47.17	2	23	0.08	23.56
SNV	C	T	No	Heterozygous	7	19	36.84	3	4	0.43	25.71
SNV	C	C	Yes	Heterozygous	12	19	63.16	5	7	0.42	24.50
SNV	T	C	No	Homozygous	14	14	100.00	13	1	0.07	24.21
SNV	A	G	No	Homozygous	14	15	93.33	10	4	0.29	23.79
SNV	T	G	No	Homozygous	7	11	63.64	1	6	0.14	25.29
SNV	A	G	No	Homozygous	8	21	38.10	7	1	0.12	25.62
SNV	A	G	No	Heterozygous	5	10	50.00	4	1	0.20	29.80
SNV	A	A	Yes	Heterozygous	5	10	50.00	4	1	0.20	26.20
SNV	C	T	No	Homozygous	56	56	100.00	41	15	0.27	31.11

Fixed Ploidy Variant Detection：結果


Rows: 231 Table view: Genome Filter

Chrom...	Region	Type	Refe...	Allele	Ref...	Len...	Zygosity	Count	Cov...	Freq...	For...	Rev...	For...	Averag...	Rea...	Rea...	# u...	# u...	Bas...	Rea...	Rea...	Hyp...	Ho...
NC_010...	1115086	SNV	A	G	No	1	Homozy...	22	24	91.67	12	10	0.45	31.82	22	24	19	19		0.99	0.95	no	No
NC_010...	1152101	SNV	G	C	No	1	Heteroz...	5	13	38.46	5	0	0.00	10.00	5	13	5	5	2.47	0.17	1.00	no	No
NC_010...	1152101	SNV	G	G	Yes	1	Heteroz...	8	13	61.54	8	0	0.00	28.38	8	13	7	7		0.38	1.00	no	No
NC_010...	1152139	SNV	A	A	Yes	1	Heteroz...	7	11	63.64	7	0	0.00	18.00	7	11	6	6		0.73	1.00	no	No
NC_010...	1152139	SNV	A	C	No	1	Heteroz...	4	11	36.36	4	0	0.00	16.25	4	11	4	4	0.00	0.45	1.00	no	No
NC_010...	1154738	SNV	A	G	No	1	Homozy...	34	34	100.00	23	11	0.32	30.53	34	34	25	25		1.00	1.00	no	No
NC_010...	1166395....	MNV	TT	CC	No	2	Homozy...	35	35	100.00	18	17	0.49	33.50	35	35	26	26		1.00	1.00	no	No
NC_010...	1178224	SNV	T	C	No	1	Homozy...	38	38	100.00	19	19	0.50	33.58	38	38	27	27		1.00	1.00	no	No
NC_010...	1186219	SNV	T	G	No	1	Heteroz...	10	27	37.04	0	10	0.00	10.40	10	27	9	9	2.63	0.02	0.15	no	No

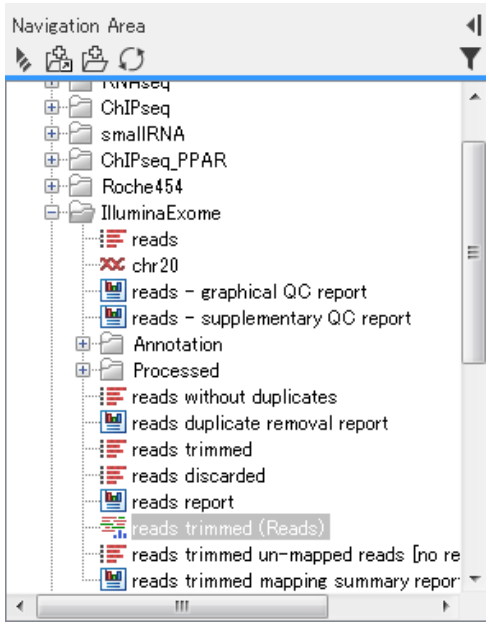


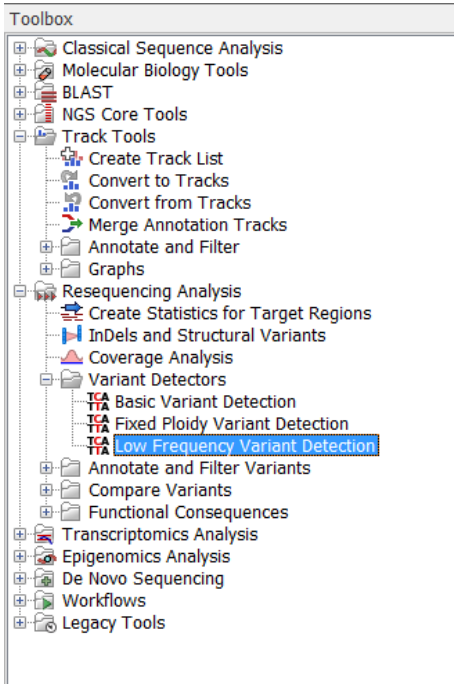
- Count: クオリティのフィルターをパスしたリードの数
- Coverage: クオリティのフィルターをパスしたリードの数
- Frequency: バリアントが見られた頻度
- Probability: バリアントのアレルの事後確率（そのアレルが尤もであるとする確率。高い方がより確度が高いという事。）
- Forward reads: その領域に見られたForwardリードの数
- Reverse reads:その領域に見られたReverseリードの数
- Forward/reverse: Forward/Total reads　または　Reverse/Total reads のうち小さい方の値。ForwardとReverseが同じなら、0.5となる。
- Average quality: 該当する領域の平均リードクオリティ。
- # unique start positions:バリアントコールに使われたリードのうちスタートポジションにあるリードの数
- # unique end positions:バリアントコールに使われたリードのうち最後の箇所にあるリードの数
- BaseQRankSum:クオリティスコアについて、参照配列と同じアレルとバアリアントのアレルについてマンホイットニーU検定を行い計算されたZスコア。これが高いほど参照配列の塩基とバリアントの塩基に差がある。
- Hyper-alleic：想定されるアレルよりも頻度が高いかどうか
- Homopolymer：ホモポリマー領域かどうか

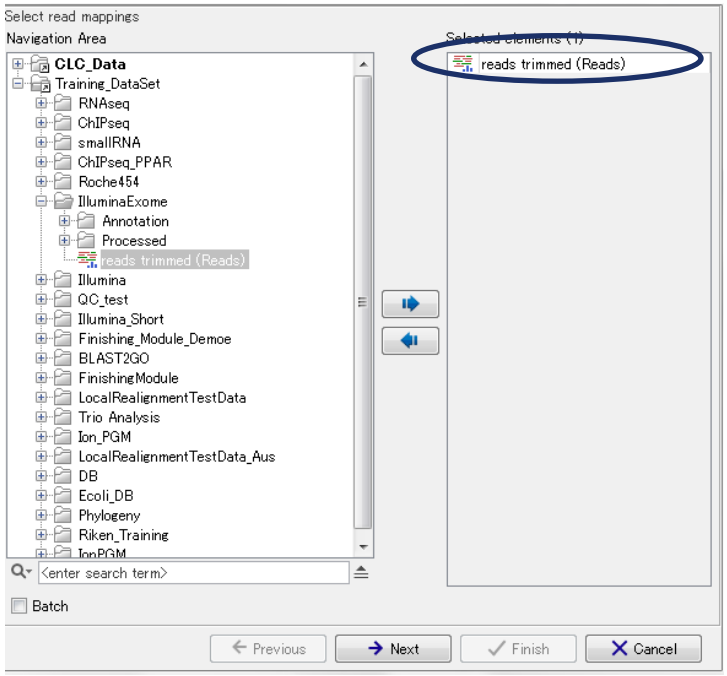
- Low frequency Variant Detection では、倍数性を仮定せず、対象となる領域が、シーケンスエラーなのか、そうではない（=バリアント）なのかを検定しています。

- Error モデルについては、Fixed Ploidy Variant Detection にて採用したエラーモデルを使い、計算し、尤度比検定を行っています。

Low Frequency Variant Detection

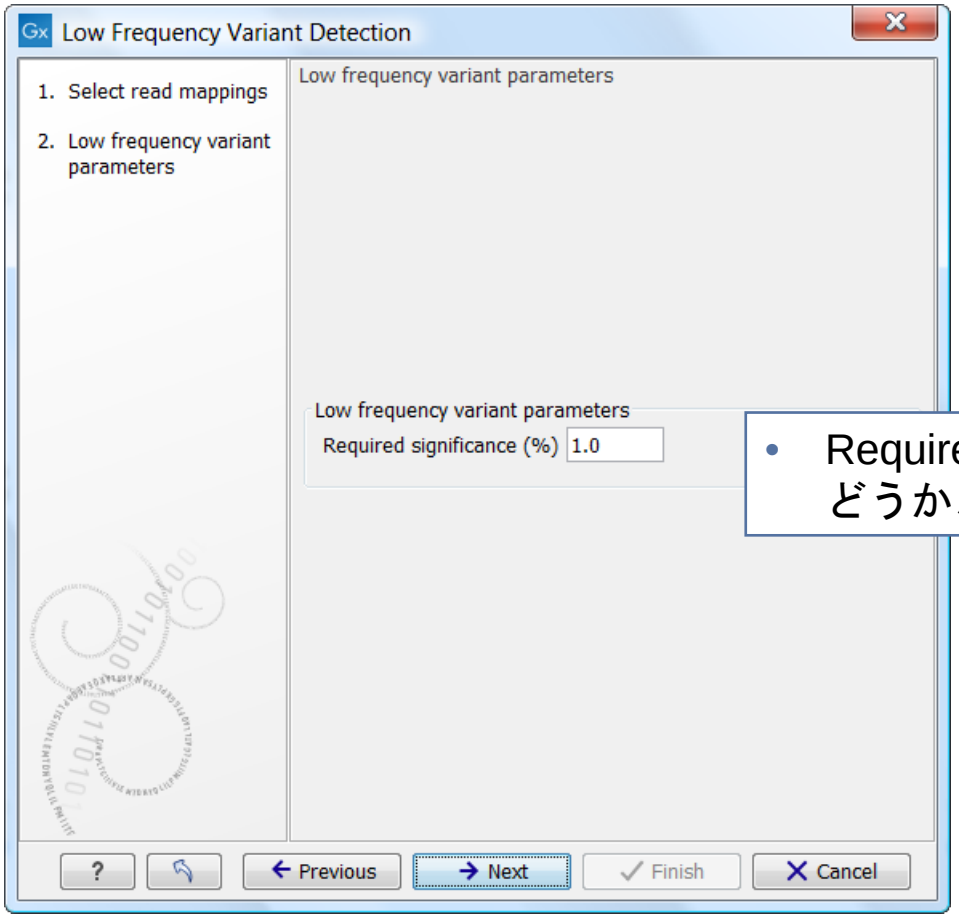




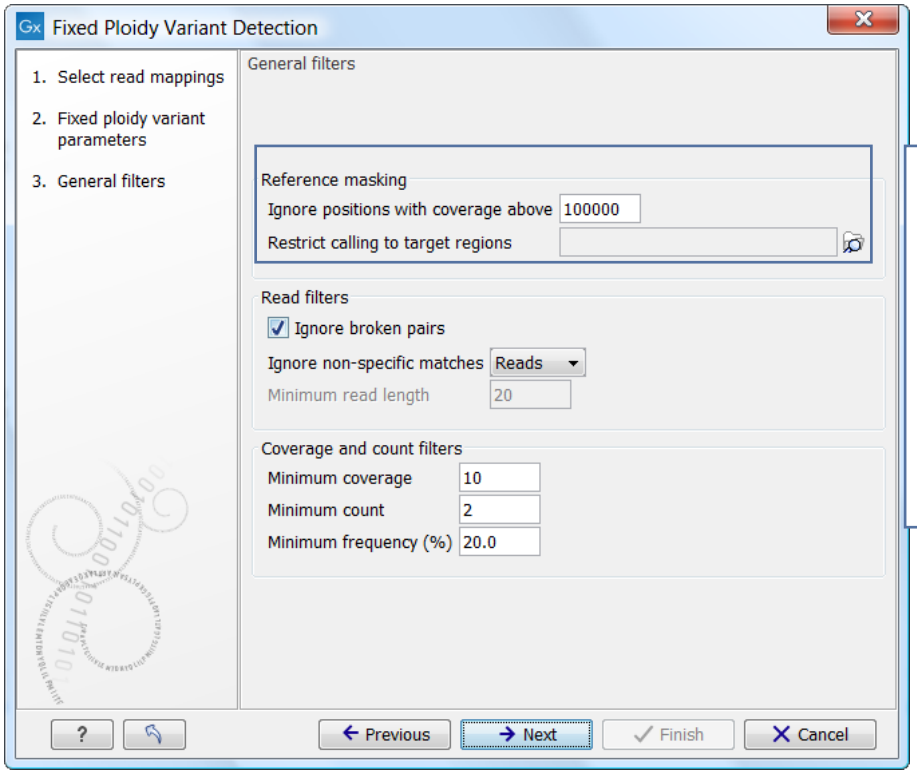


- Navigation Areaからマッピングデータを選択。
- Toolboxから Resequencing Analysis > Variant Detectors > Low Frequency Variant Detection を選択、ダブルクリック。
- ウィザードが起動し、選択したデータが選ばれていることを確認。

Low Frequency Variant Detection

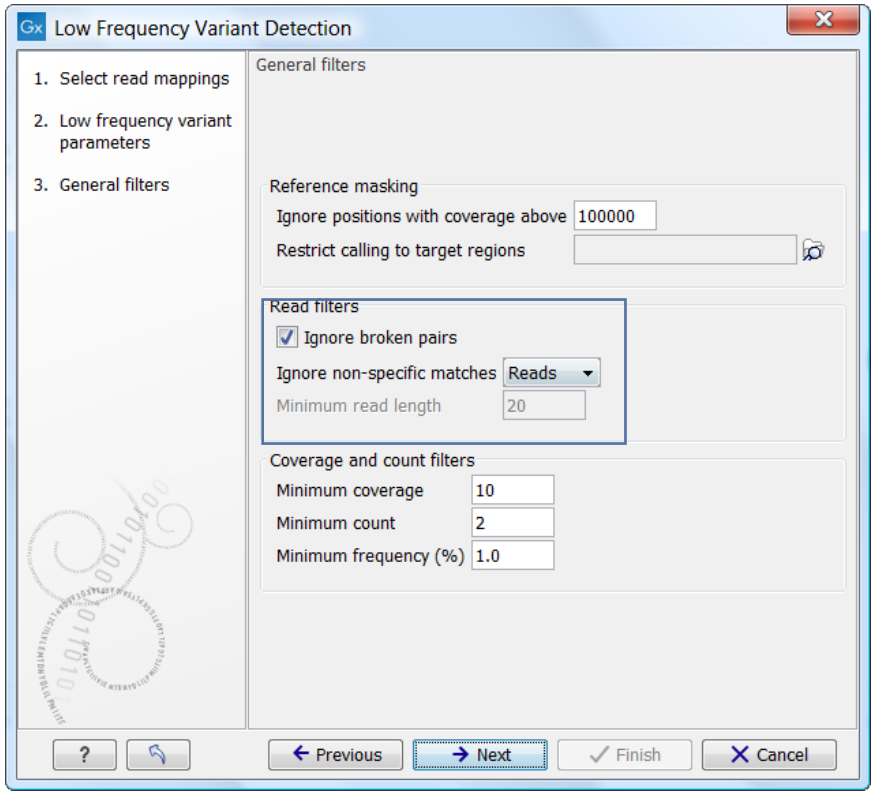


- Required significance :シーケンスエラーかどうか、検定の際の閾値



Reference masking

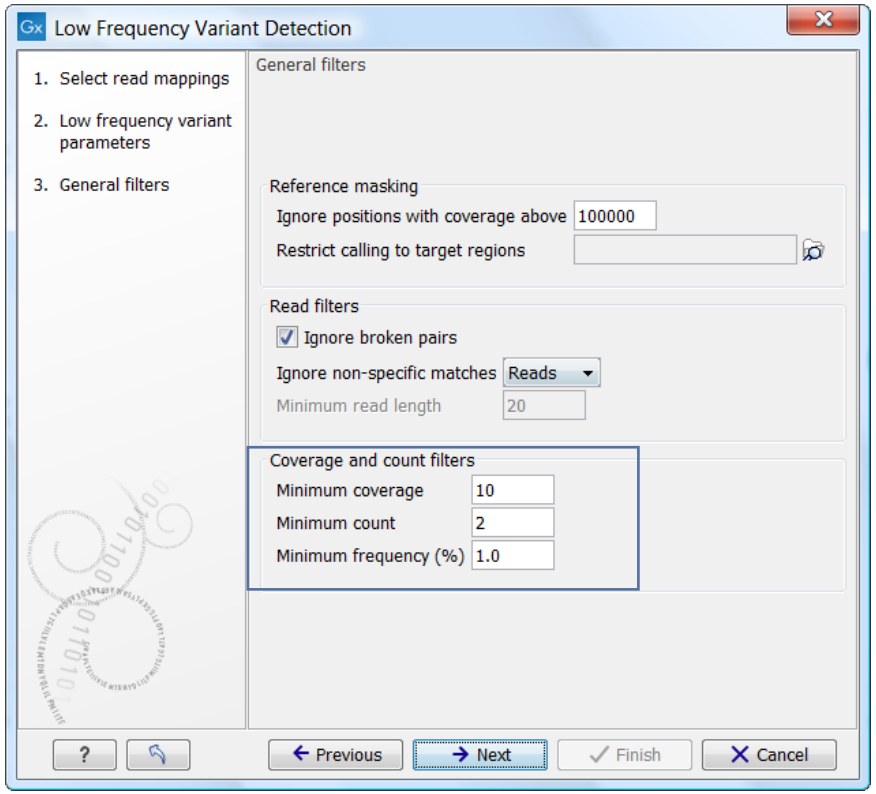
- Ignore positions with coverage above：カバレッジが指定した数字以上のバリアントについてリストに含めない
- Restrict calling to target regions：バリアントを検出したい領域の指定（アノテーショントラックで指定）



Read filters

- Ignore broken pairs：ペアエンドのリードでペアと認識されなかったリードをバリアント検出の計算に含めるかどうか
- Ignore non-specific matches：「Reads」を選択すると、non-specificなマッチのリードを計算に含めなくなり、「Regions」を選択すると、1本でもnon-specificなリードが含まれる場合、その領域のバリアントを検出しません。
- Minimum read length：Ignore broken pairとIgnore non-specific regions が指定された場合、このフィルターの対象となる最小のリードの長さの設定が必要です。これは非常に短いリードは、その短さからnon-specificになる可能性があるためです。

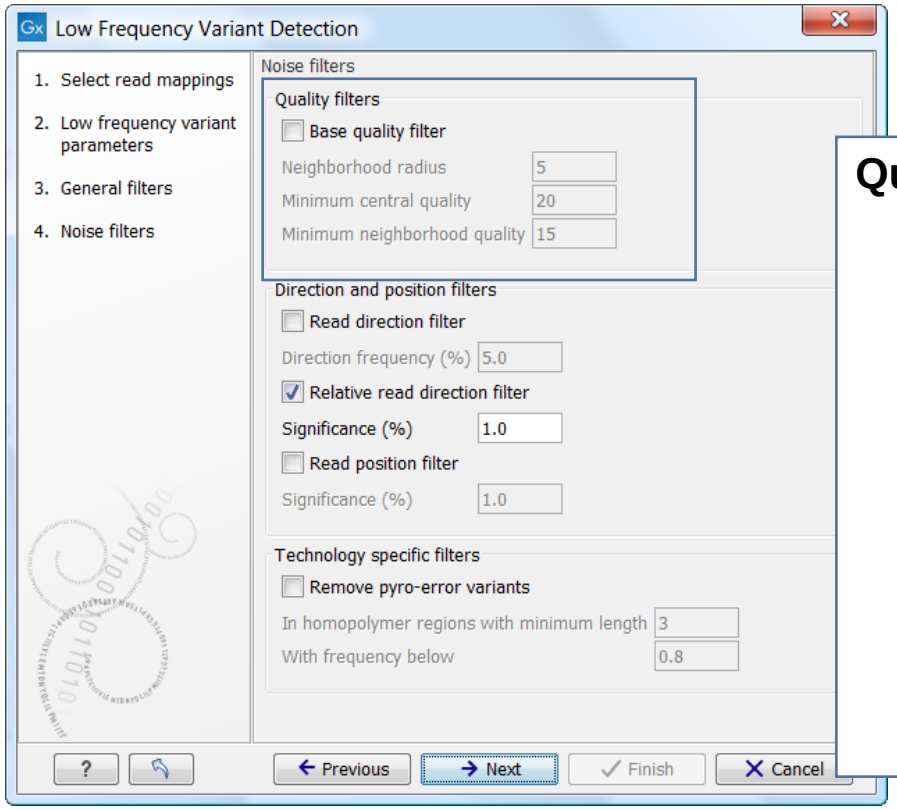
Low Frequency Variant Detection



Coverage and count filters

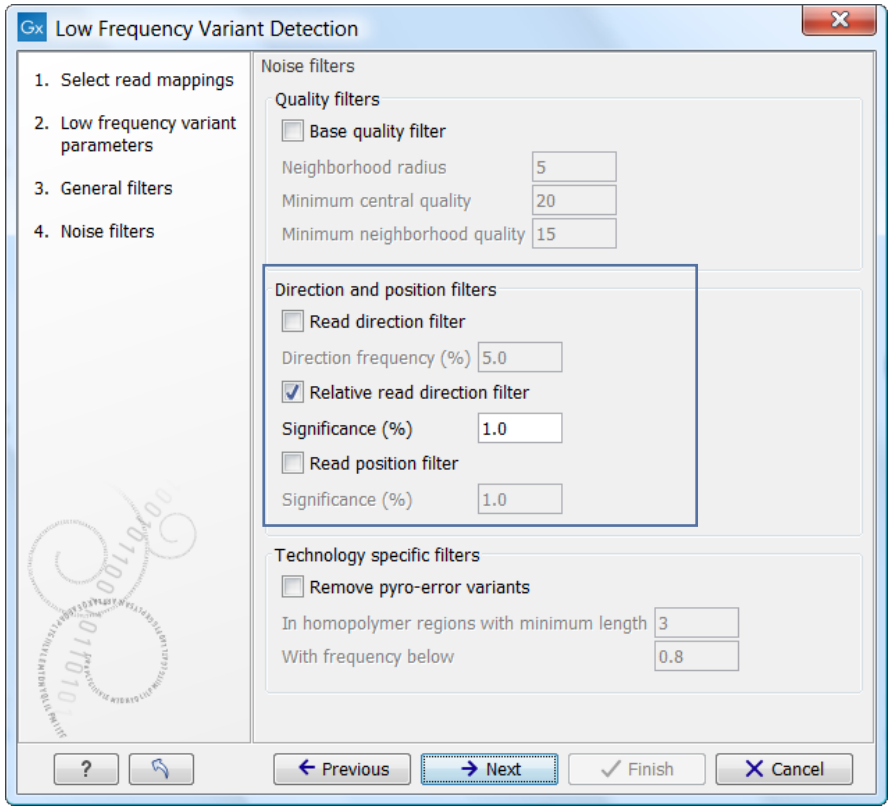
Minimum coverage：最小カバレッジ
Minimum count：バリアントを支持するリードの最低カウント数
Minimum frequency (%)：最小頻度

Low Frequency Variant Detection



Quality filter

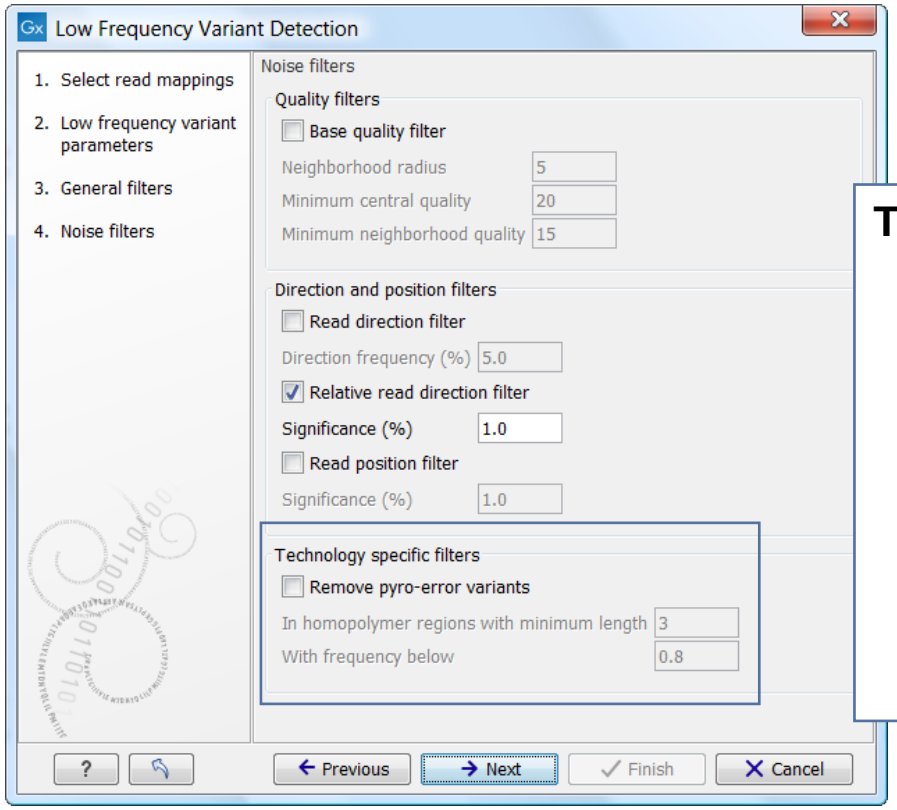
- Base quality filter:塩基のクオリティに関するフィルター
  - Neighborhood radius：クオリティフィルターの対象とする横方向の塩基数（奇数）
  - Minimum central quality：縦方向の数（リード数）
  - Minimum neighborhood quality：Neighborhood radiusで指定した範囲の最低クオリティ（Phred score）



Direction and position filters：

- Read direction filter：どちらか一方の方向のリードが多数見られる場合にそれを排除（ただし、アンプリコンには適していません）。
  - Relative read direction filter：リードの方向が一方のみに偏りすぎていないか、全体のForwardとReverseのバランスを見て統計検定を行う。Significanceで閾値を入力。
  - Read position filter：システマティックなエラーを取り除くために用いるツールでハイブリダイゼーションを行った場合のデータに有効。リードを5つのセグメントに分割し、バリアントの見られるポジションの５つのセグメントに分割されたリードの分布が全体のそれと似ているかどうか検定を行う。Significanceで閾値を入力。

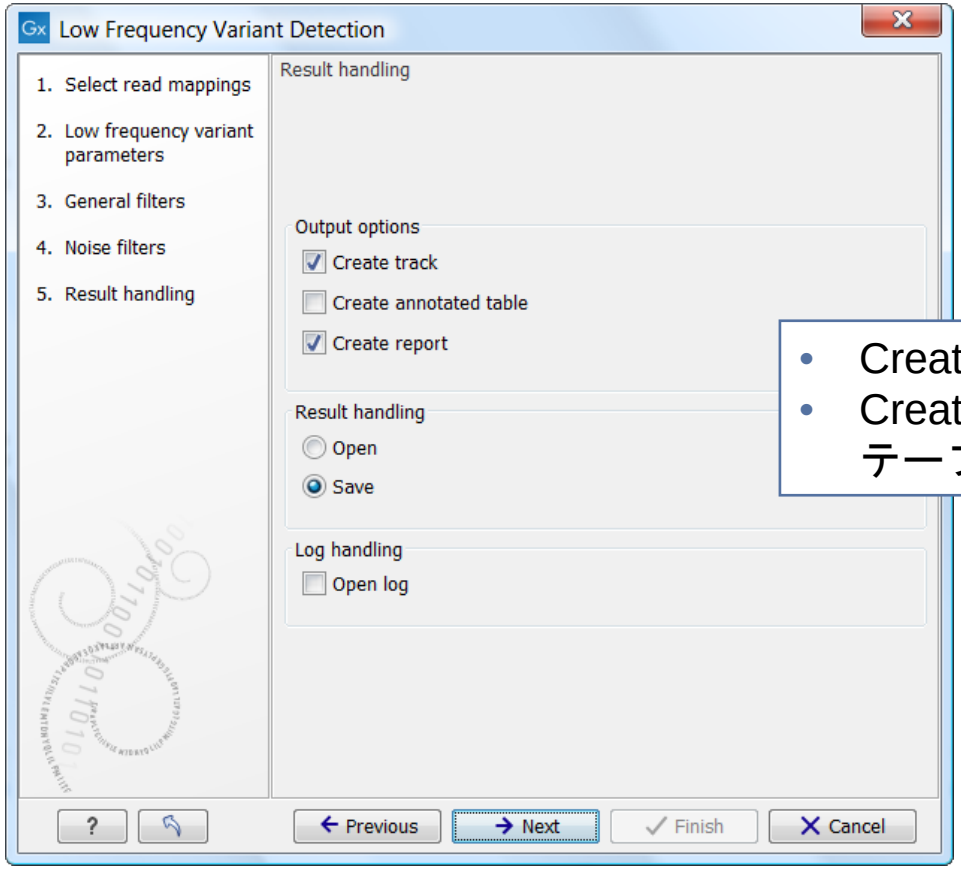
Low Frequency Variant Detection



Technology specific filters

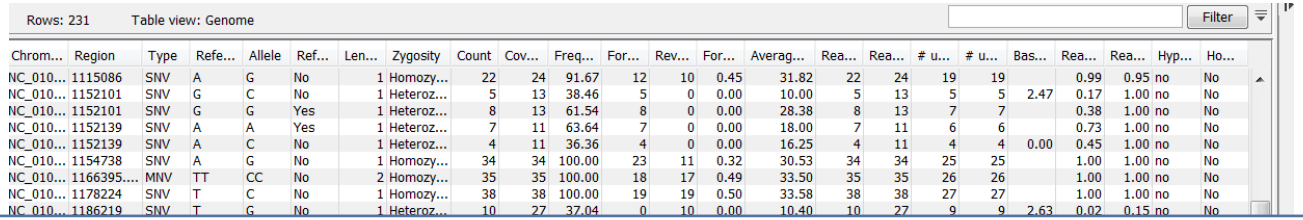
- Remove pyro-error variants：ホモポリマー領域に対するエラーの除去
  - In homopolymer regions with minimum length：指定した長さのホモポリマー領域のInDelを取り除く。
  - With frequency below：指定した頻度以下のものについてのみフィルターを適用。

Low Frequency Variant Detection

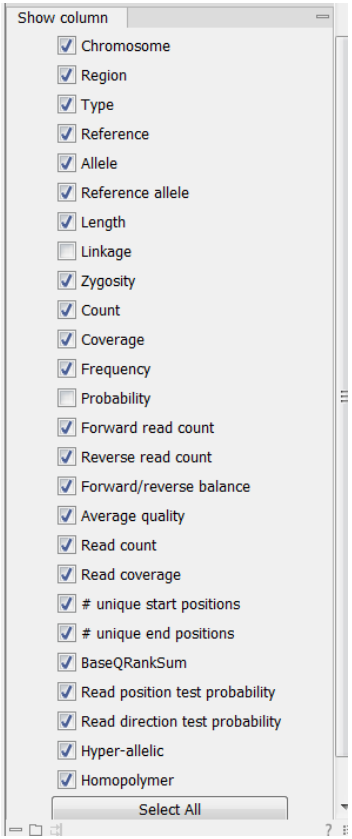


- Create track: トラックの作成
- Create annotated table: アノテーション付のテーブルの作成

Low Frequency Variant Detection：結果



Chrom...	Region	Type	Refe...	Allele	Ref...	Len...	Zygosity	Count	Cov...	Freq...	For...	Rev...	For...	Averag...	Rea...	Rea...	# u...	# u...	Bas...	Rea...	Rea...	Hyp...	Ho...
NC_010...	1115086	SNV	A	G	No	1	Homozy...	22	24	91.67	12	10	0.45	31.82	22	24	19	19		0.99	0.95	no	No
NC_010...	1152101	SNV	G	C	No	1	Heteroz...	5	13	38.46	5	0	0.00	10.00	5	13	5	5	2.47	0.17	1.00	no	No
NC_010...	1152101	SNV	G	G	Yes	1	Heteroz...	8	13	61.54	8	0	0.00	28.38	8	13	7	7		0.38	1.00	no	No
NC_010...	1152139	SNV	A	A	Yes	1	Heteroz...	7	11	63.64	7	0	0.00	18.00	7	11	6	6		0.73	1.00	no	No
NC_010...	1152139	SNV	A	C	No	1	Heteroz...	4	11	36.36	4	0	0.00	16.25	4	11	4	4	0.00	0.45	1.00	no	No
NC_010...	1154738	SNV	A	G	No	1	Homozy...	34	34	100.00	23	11	0.32	30.53	34	34	25	25		1.00	1.00	no	No
NC_010...	1166395....	MNV	TT	CC	No	2	Homozy...	35	35	100.00	18	17	0.49	33.50	35	35	26	26		1.00	1.00	no	No
NC_010...	1178224	SNV	T	C	No	1	Homozy...	38	38	100.00	19	19	0.50	33.58	38	38	27	27		1.00	1.00	no	No
NC_010...	1186219	SNV	T	G	No	1	Heteroz...	10	27	37.04	0	10	0.00	10.40	10	27	9	9	2.63	0.02	0.15	no	No



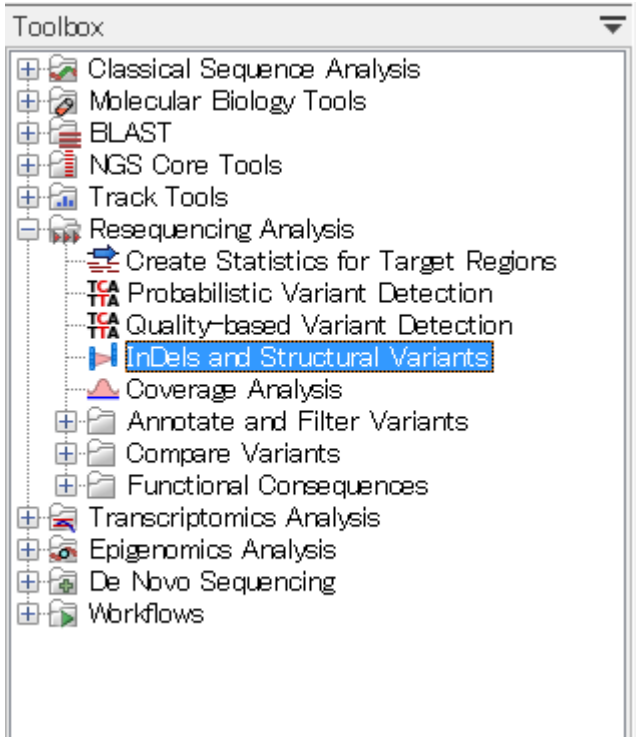
- Count: クオリティのフィルターをパスしたリードの数
- Coverage: クオリティのフィルターをパスしたリードの数
- Frequency: バリアントが見られた頻度
- Probability: バリアントのアレルの事後確率（そのアレルが尤もであるとする確率。高い方がより確度が高いという事。）
- Forward reads: その領域に見られたForwardリードの数
- Reverse reads:その領域に見られたReverseリードの数
- Forward/reverse: Forward/Total reads　または　Reverse/Total reads のうち小さい方の値。ForwardとReverseが同じなら、0.5となる。
- Average quality: 該当する領域の平均リードクオリティ。
- # unique start positions:バリアントコールに使われたリードのうちスタートポジションにあるリードの数
- # unique end positions:バリアントコールに使われたリードのうち最後の箇所にあるリードの数
- BaseQRankSum:クオリティスコアについて、参照配列と同じアレルとバアリアントのアレルについてマンホイットニーU検定を行い計算されたZスコア。これが高いほど参照配列の塩基とバリアントの塩基に差がある。
- Hyper-alleic：想定されるアレルよりも頻度が高いかどうか
- Homopolymer：ホモポリマー領域かどうか

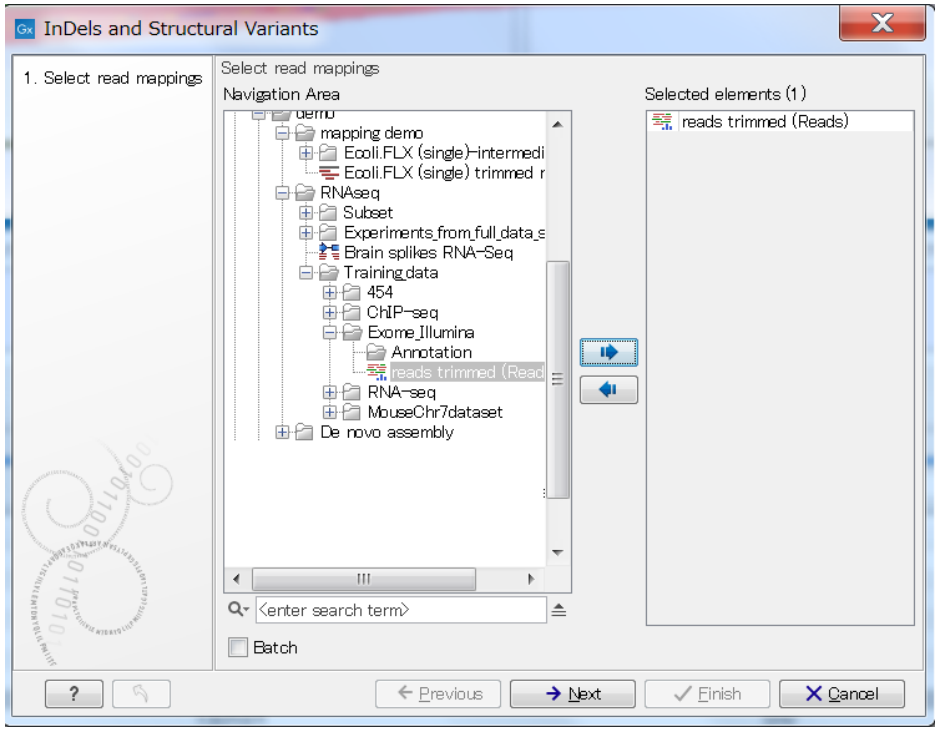
挿入、欠失と構造変異解析

- Quality Based Variant Detection やProbabilistic Variant Detection では変異やInDelを検出できました。
- しかしながら大きなInDelの検出や構造変異については、上記ツールでの検出は難しい場合があります。

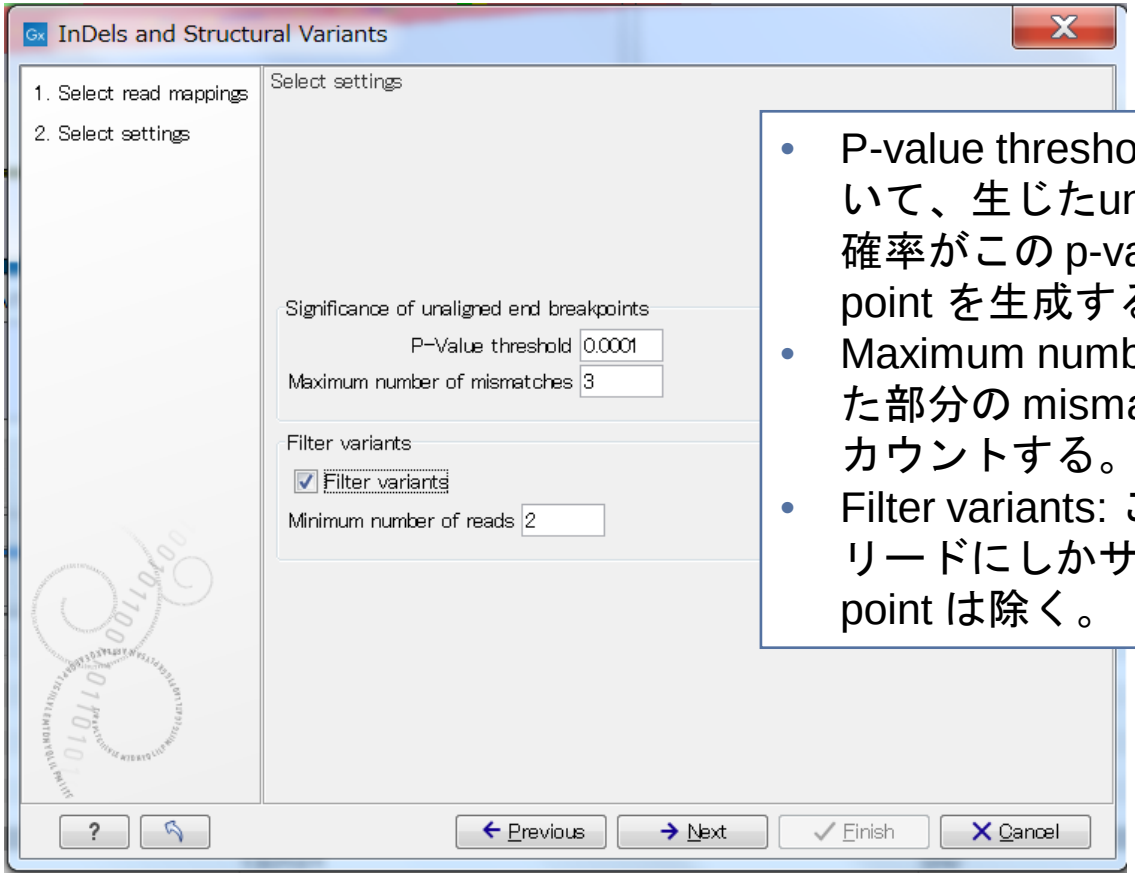
- アルゴリズムにとっては、大きなInsertionやDeletionを受け入れるよりは、Unaligned end とするほうがスコアを大きくできるからです。
- InDel and Structural Variants ツールでは、このUnaligned end に着目して、大きなInDelや構造変異を見つけます。
- Unaligned end が別の領域に十分な量マップすることができれば、そこまでの距離のInsertionやDeletion、構造変異と考えられます。
- 注意：このツールでは、同一染色体内の構造変異のみが検出可能です。

InDels and Structural Variants

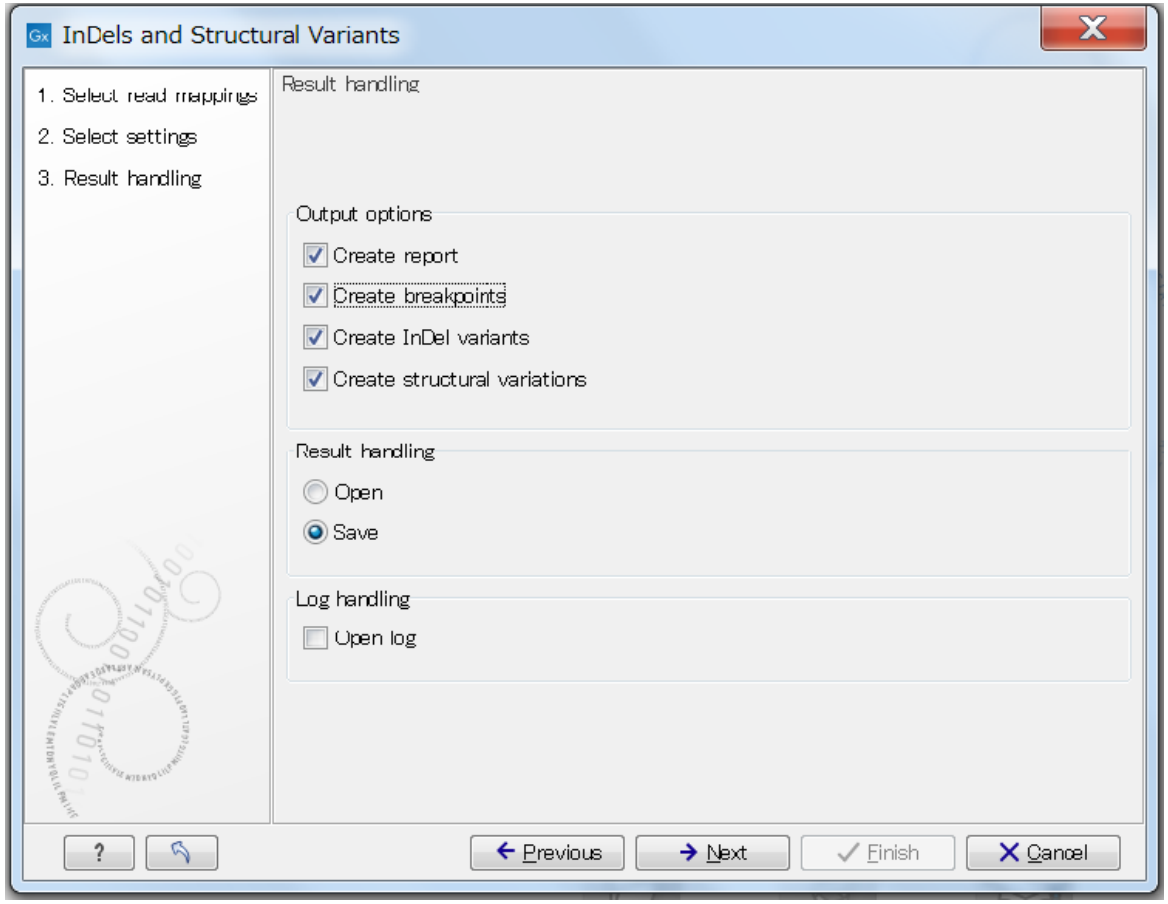




- Navigation Areaからマップするデータを選択。
- Toolboxから InDels and Structural Variantsを選択、ダブルクリック。
- ウィザードが起動し、選択したデータが選ばれていることを確認。

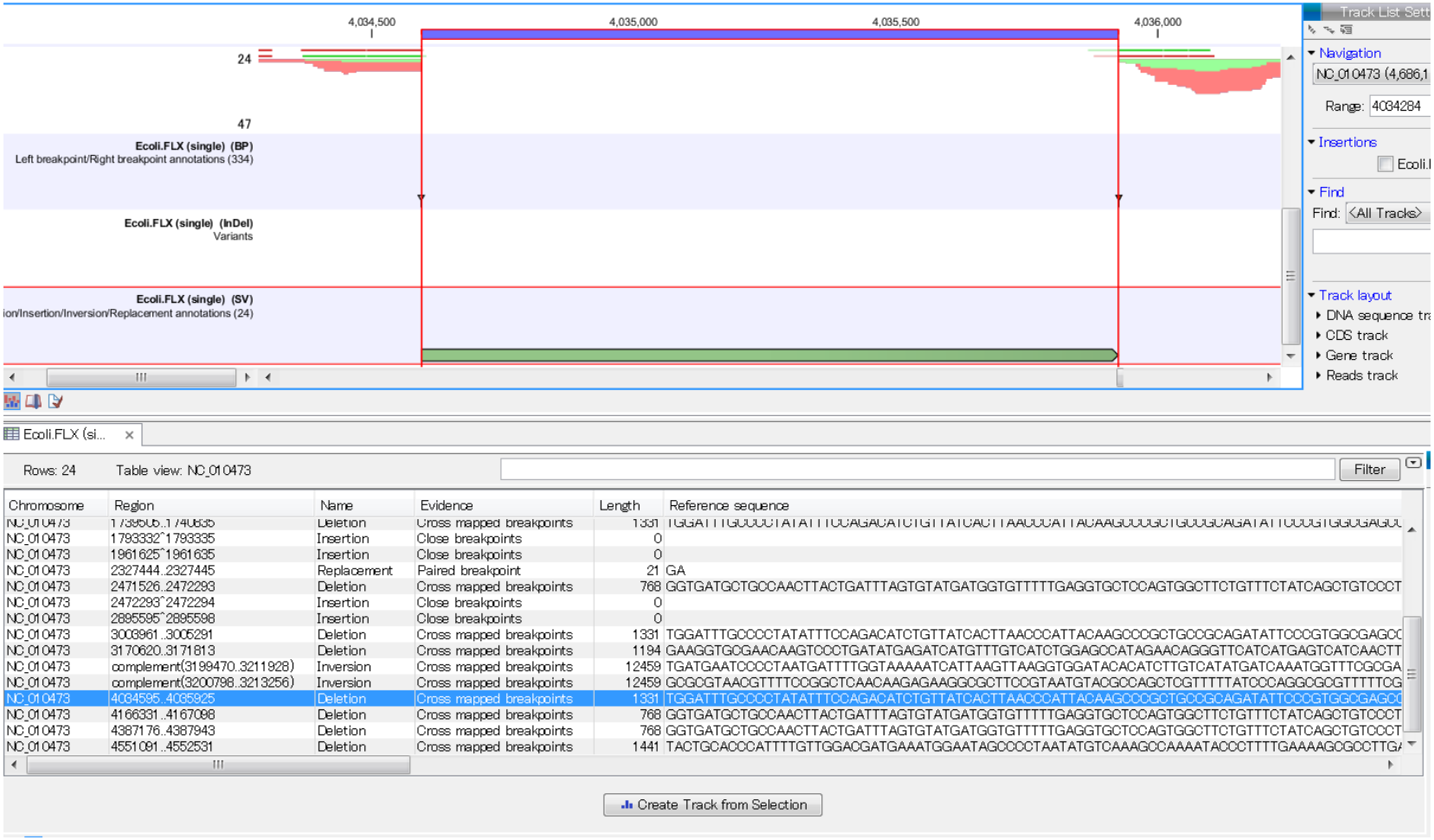


- P-value threshold: 得られた coverage 数において、生じたunaligned end の数が得られる確率がこの p-valueより少ない場合、break point を生成する。
- Maximum number of mismatches: align された部分の mismatch がこの数以下のもののみカウントする。
- Filter variants: ここで指定した数より少ないリードにしかサポートされていない break point は除く。

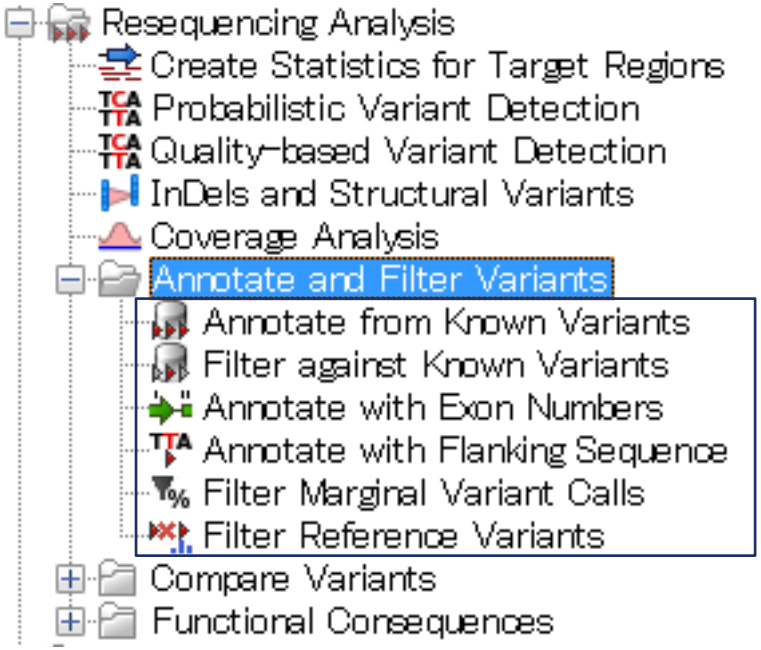


- Output について設定する。

InDels and Structural Variants

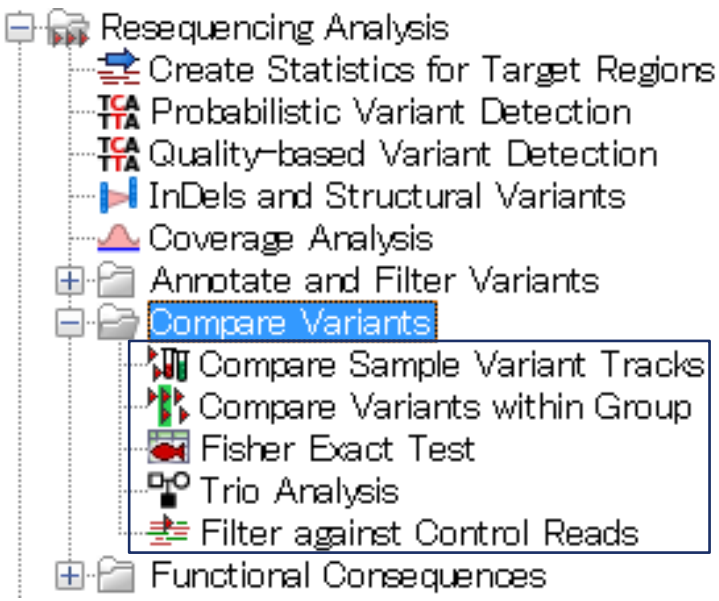


リシーケンシング解析：その他の機能



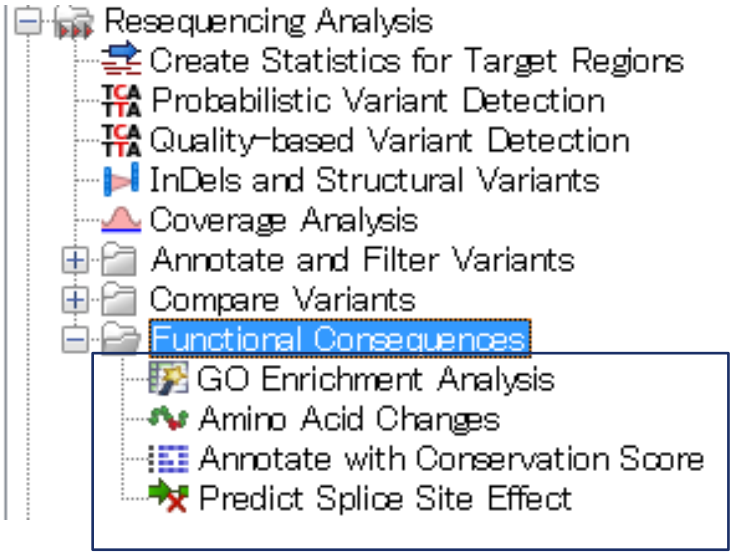
- Annotate from Known Variants : known variants とオーバーラップする variants にアノテーション付けする
- Filter against Known variants : known variants と比較してフィルタリングする
- Annotate with Exon Numbers : exon の番号をアノテーションに追加する
- Annotate with Flanking Sequences : reference の隣接する塩基とともにアノテーション付けする
- Filter Marginal Variant Calls : Variant frequency, Forward/reverse balance, Average base quality などの条件でフィルタリングする
- Filter Reference Variants : reference allele variants をフィルタリングする

リシーケンシング解析：その他の機能



- Compare Sample Variant Tracks: 2つの variant track を比較して、共通する、または、異なる variant を出力する (Ver6.5で追加)。
- Compare Variants within Group : グループの中で common variants を検索する。Frequency を % で指定できる。
- Fisher Exact Test : Case-control study で、case に有意に存在する variants を検出する。
- Trio Analysis : 子供と両親のデータを用いて trio 解析を行う。Variants が親に由来するのか、de novo なのかをレポートする。
- Filter against Control Reads : Control に存在する variant をフィルタリングする。

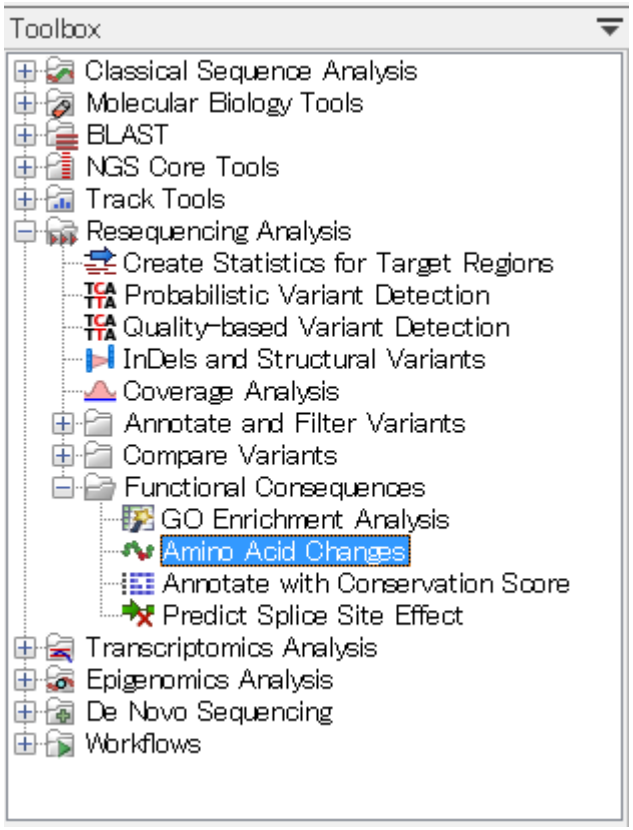
リシーケンシング解析：その他の機能

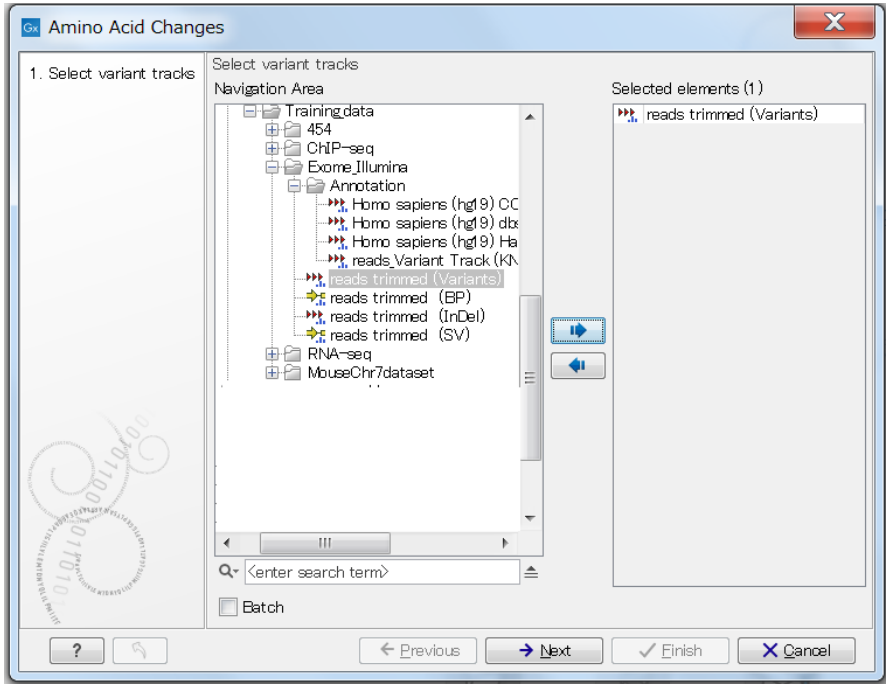


- GO Enrichment Analysis : 検出された variants が含まれる遺伝子にどのような Gene Ontology と関連するものが多いのかを解析する。
- Amino Acid Changes : variants に、アミノ酸置換に関するアノテーション付けを行う。
- Annotate with Conservation Score : 異なる種におけるアミノ酸の保存の程度に関する情報をアノテーション付けする。保存の度合いが高いほど、機能的に重要であると期待される。
- Predict Splice Site Effect : variants の splice site に対する影響を予測する。

アミノ酸置換の解析

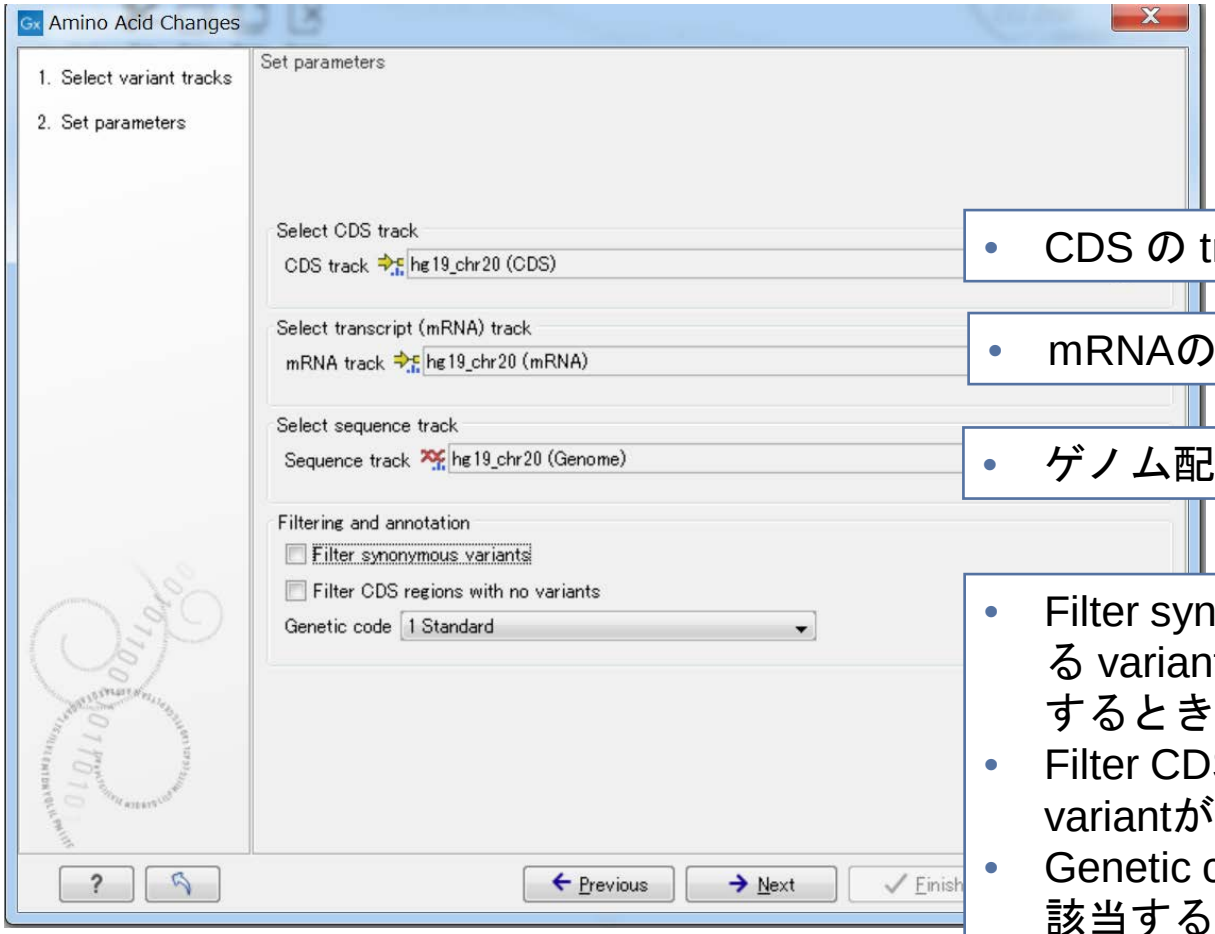
アミノ酸置換の解析





- Resequencing Analysis – Amino Acid Changes を選択
- Variant データを選択

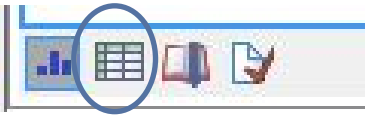
アミノ酸置換の解析

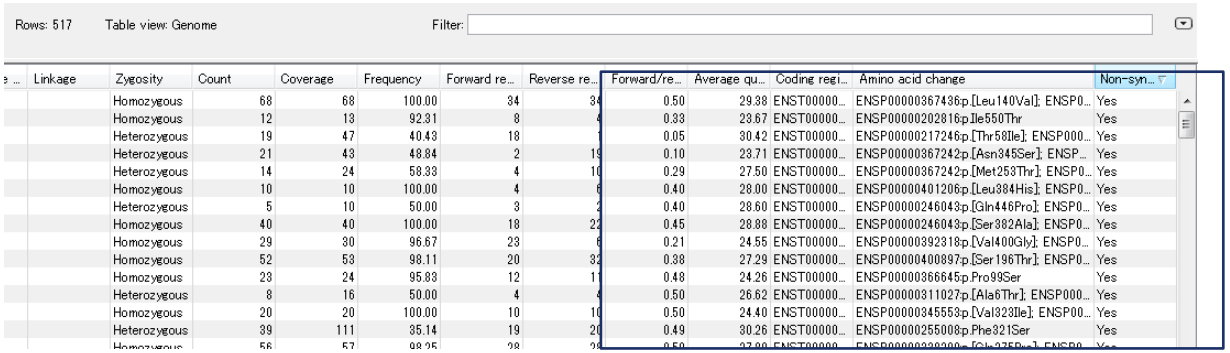


- CDS の track を選択
- mRNAのtrackを選択
- ゲノム配列のtrackを選択
- Filter synonymous: アミノ酸が変化する variant のみをアノテーション付けするときにチェックする
- Filter CDS regions with no variants: variantが無い領域をフィルターする
- Genetic code: 解析しているゲノムで該当するものを選択する

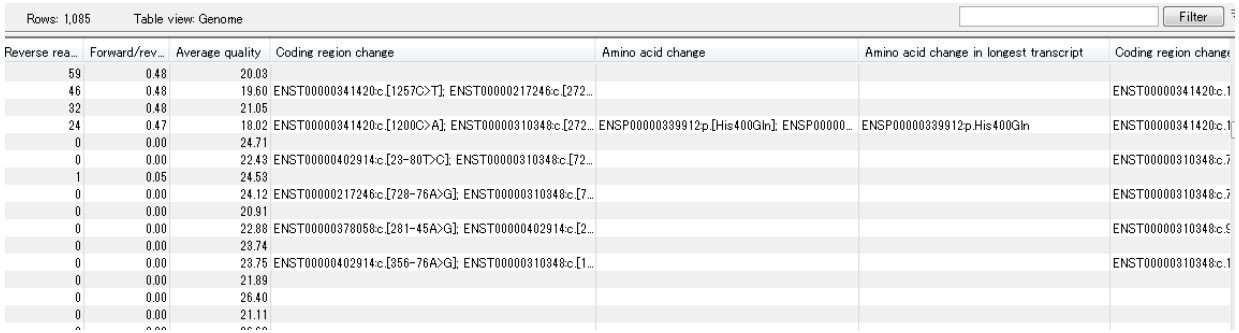
アミノ酸置換の解析

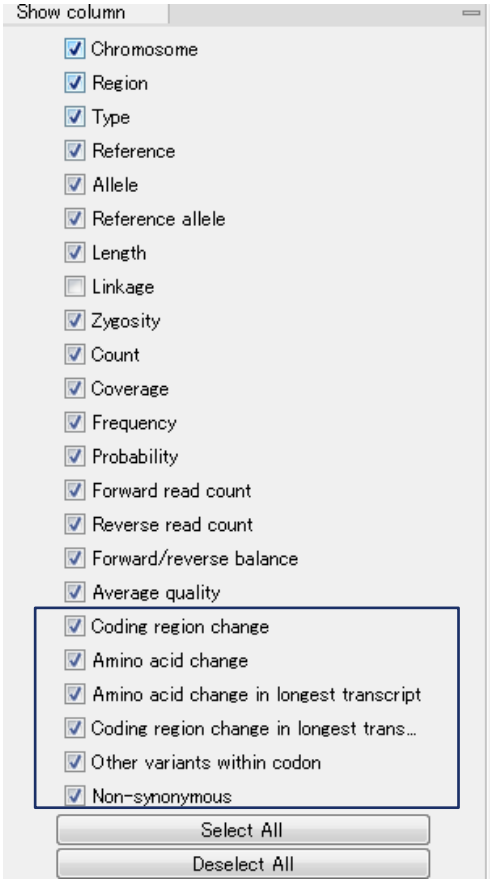
Variant Track





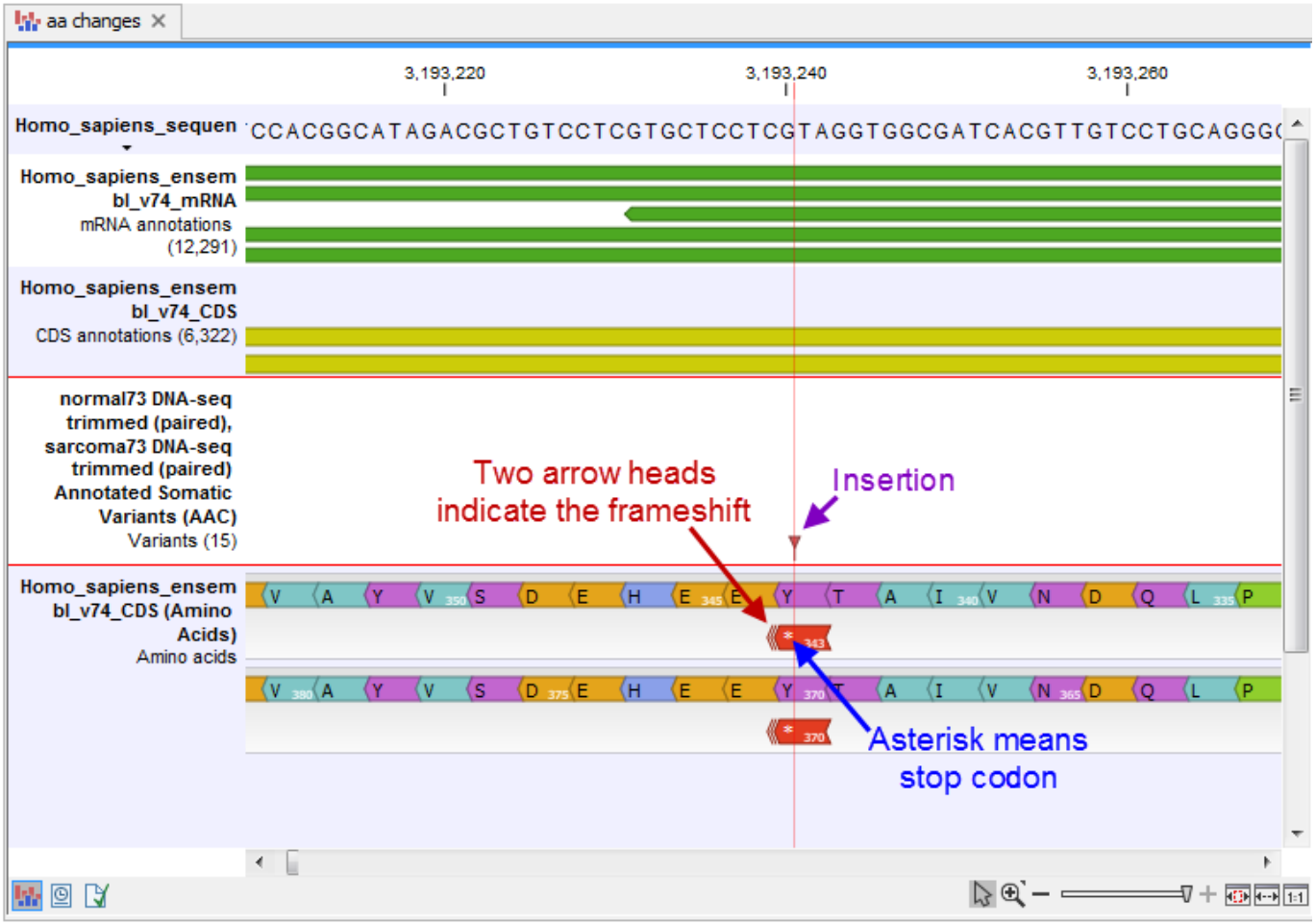
アミノ酸置換の解析





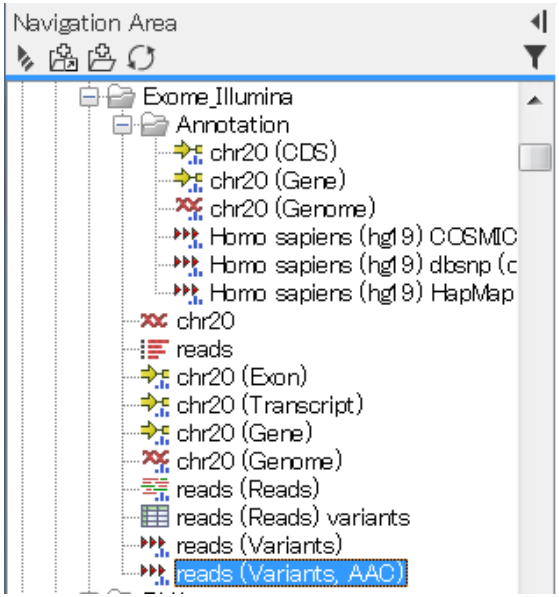
- アミノ酸置換に関する情報がテーブルに追加されました

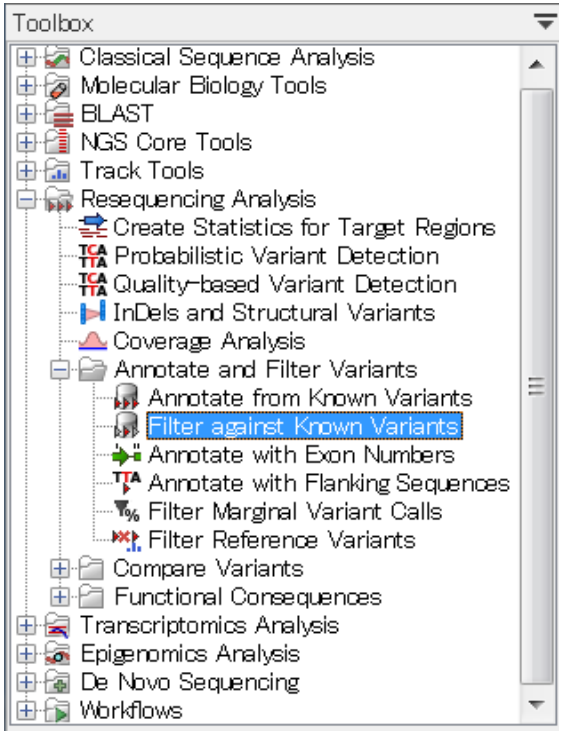
Amino Acid Track

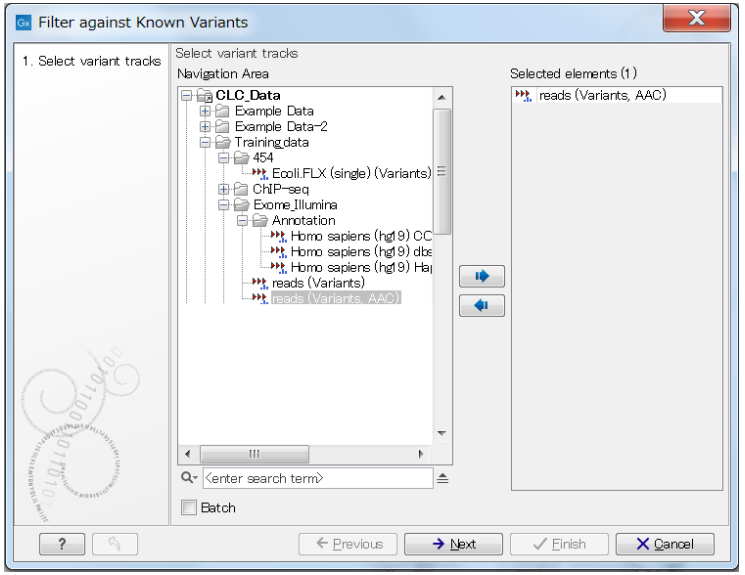


変異のフィルタリング

既知の variants の filter：Filter against Known Variants

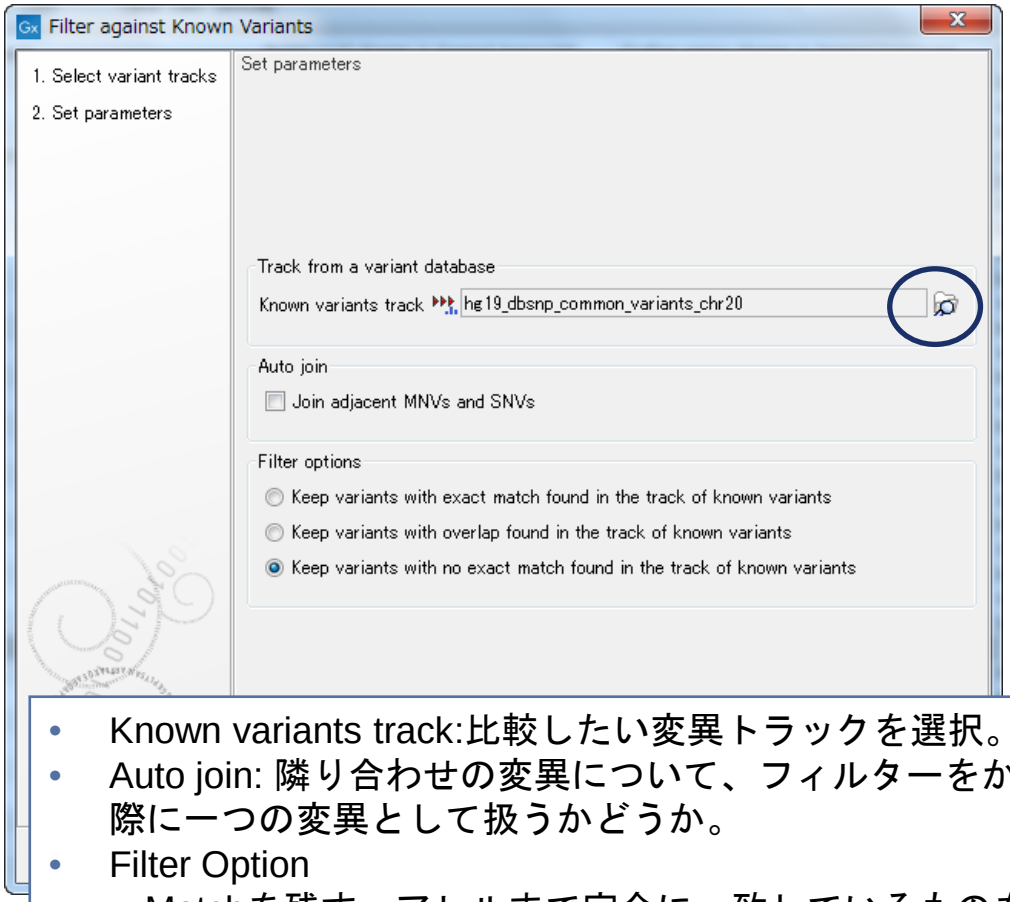


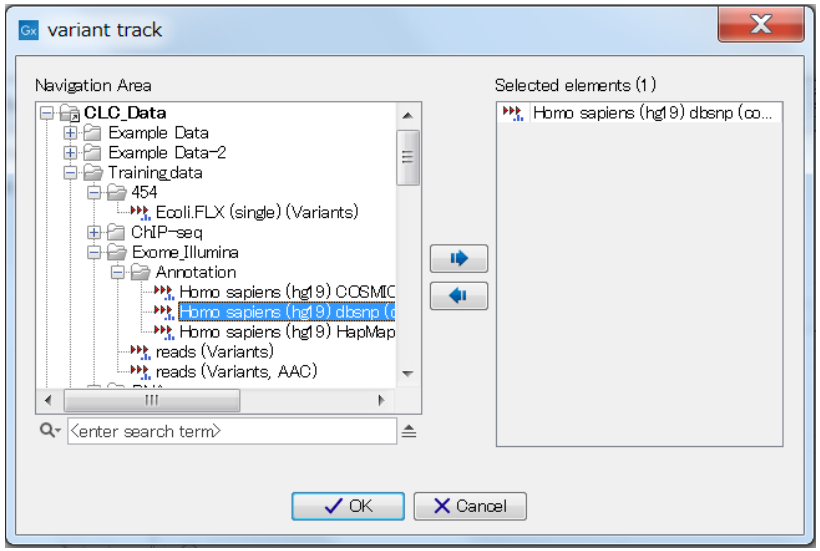




- Navigation Areaから変異トラック(アミノ酸置換を調べたもの)を選択。
- Toolboxから Resequencing Analysis > Annotate and Filter Variants > Filter against Known Variants を選択、ダブルクリック。
- ウィザードが起動し、選択したデータが選ばれていることを確認。

既知の variants の filter：Filter against Known Variants

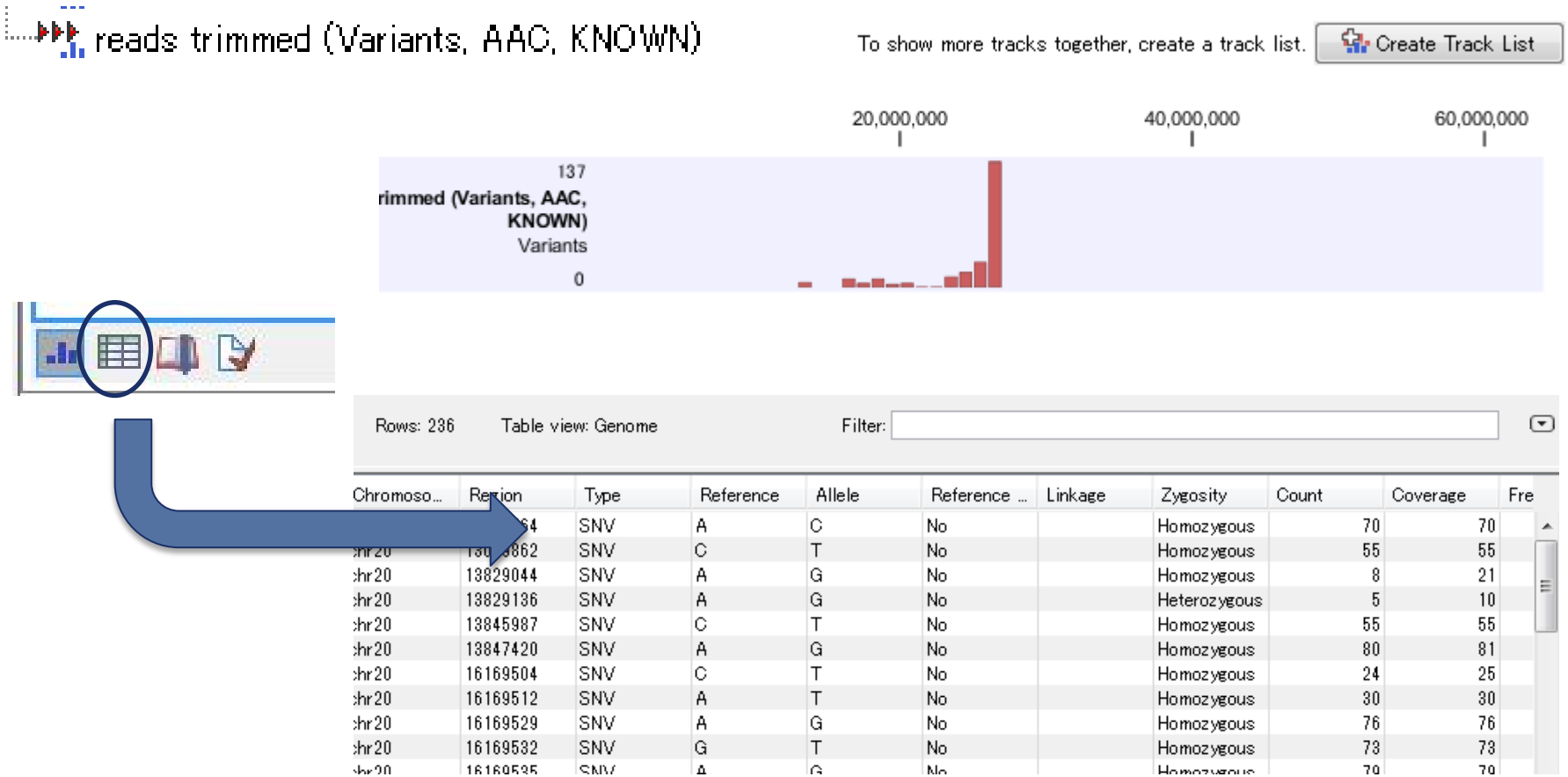




- Known variants track:比較したい変異トラックを選択。
- Auto join: 隣り合わせの変異について、フィルターをかける際に一つの変異として扱うかどうか。
- Filter Option
  Matchを残す：アレルまで完全に一致しているものを残す。
  Overlapを残す：オーバーラップがあるものを残す。
  Not matchを残す：完全に一致しなかったものを残す。

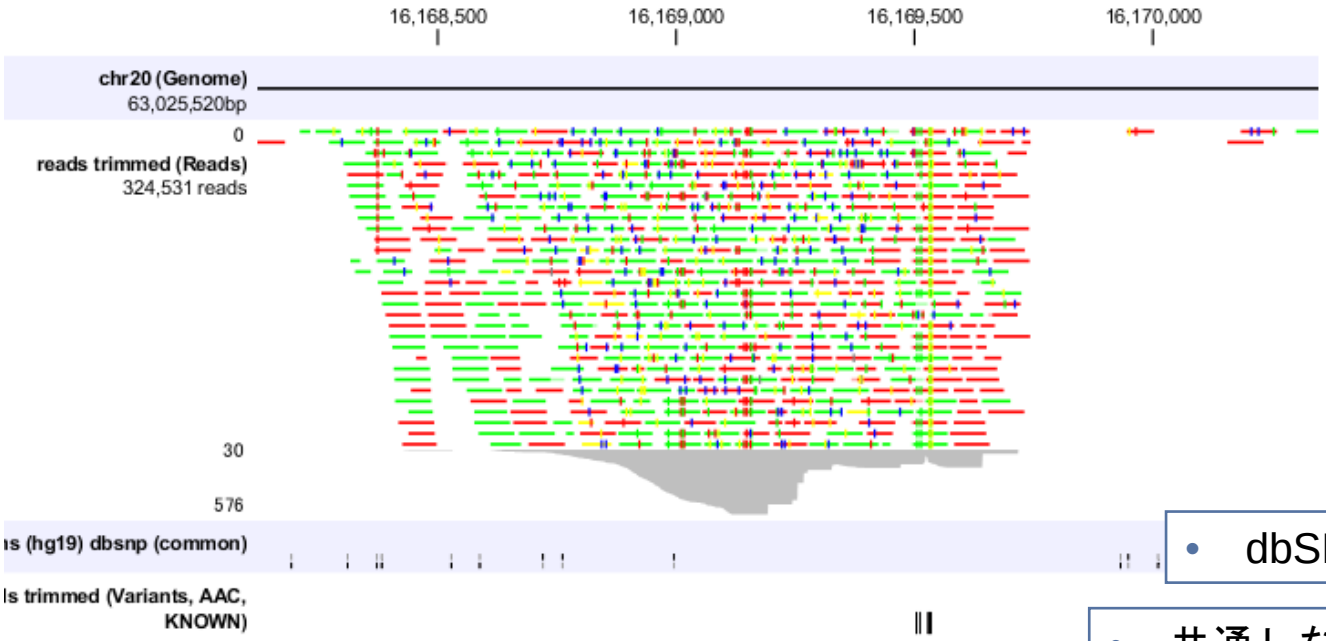
- SNPのトラックを選択。

既知の variants の filter：Filter against Known Variants



既知の variants の filter：Filter against Known Variants

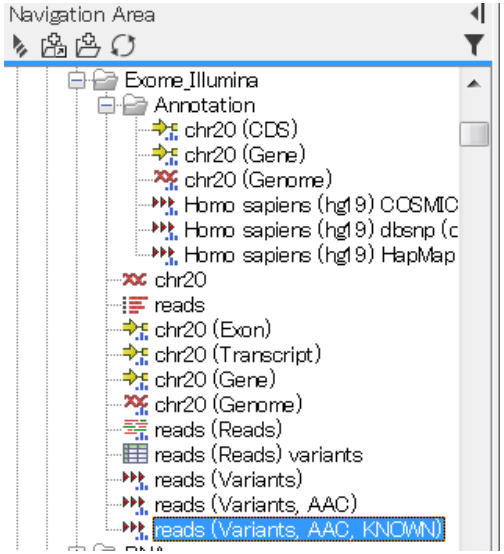
トラックで表示

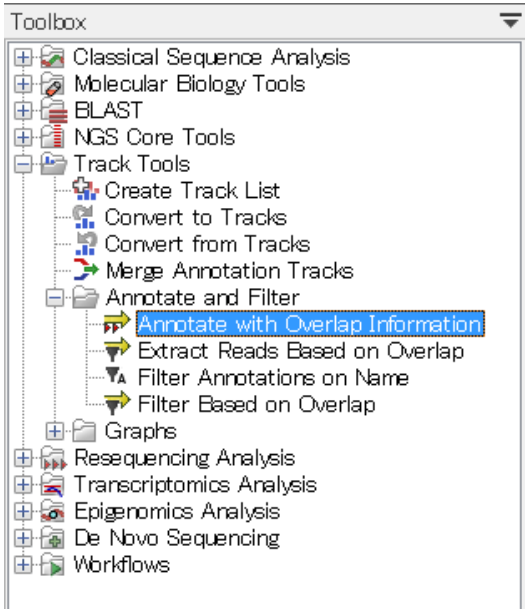


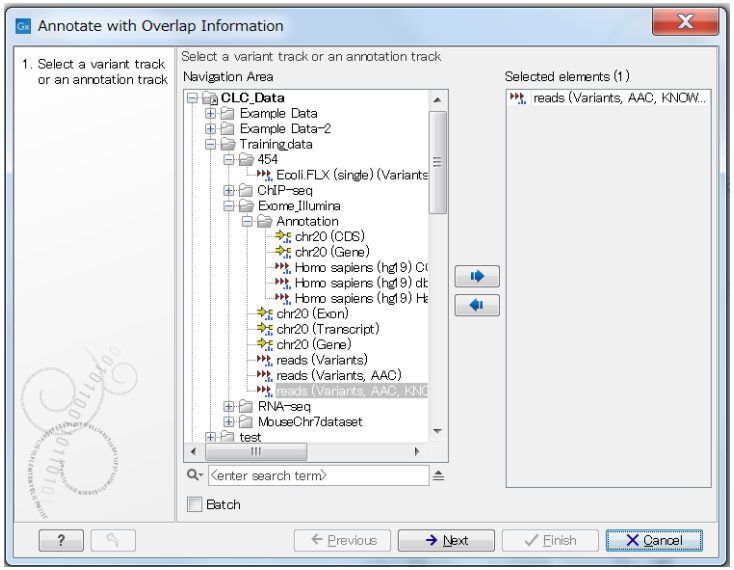
- dbSNP
- 共通しなかったもの

遺伝子のアノテーション

アノテーション付け：Annotate with Overlap Information

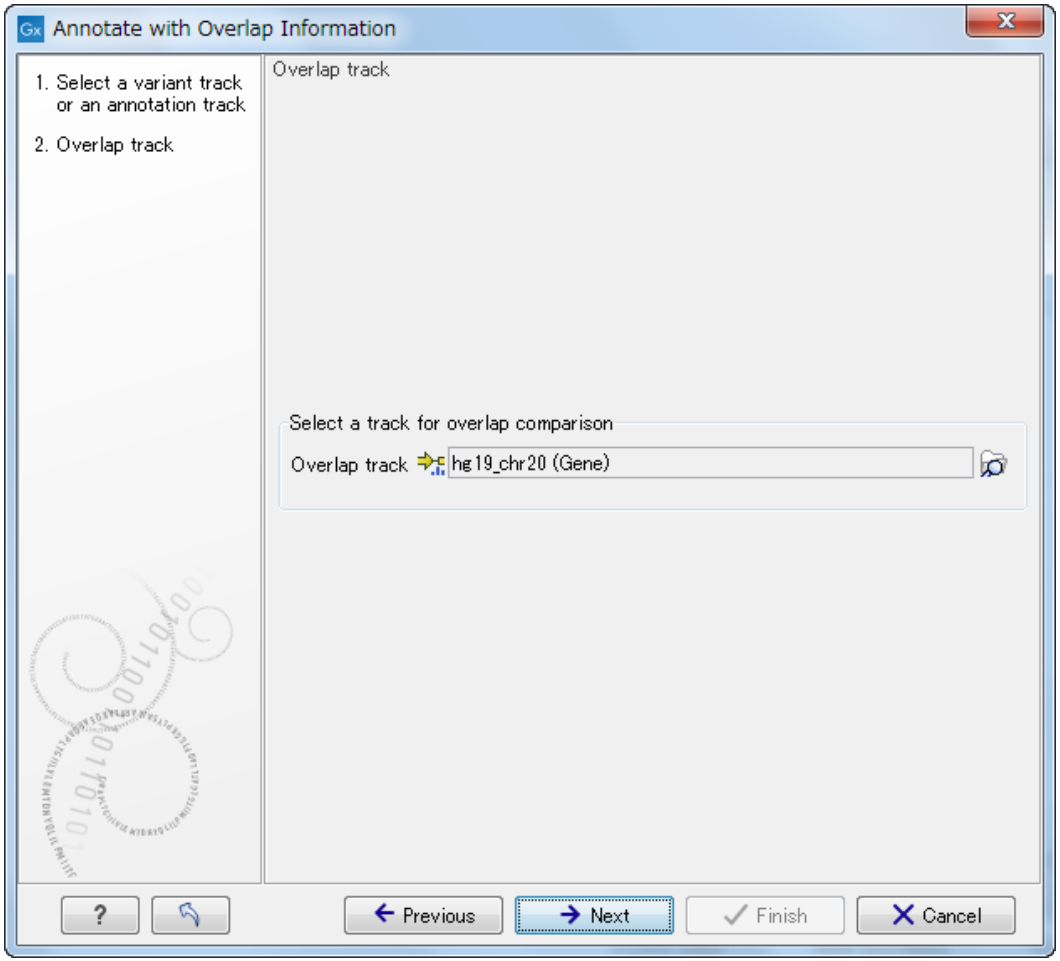






- Navigation Areaから変異トラック(フィルタリングしたもの)を選択。
- Toolboxから Track tools > Annotate and Filter > Annotate with Overlap Information を選択、ダブルクリック。
- ウィザードが起動し、選択したデータが選ばれていることを確認。

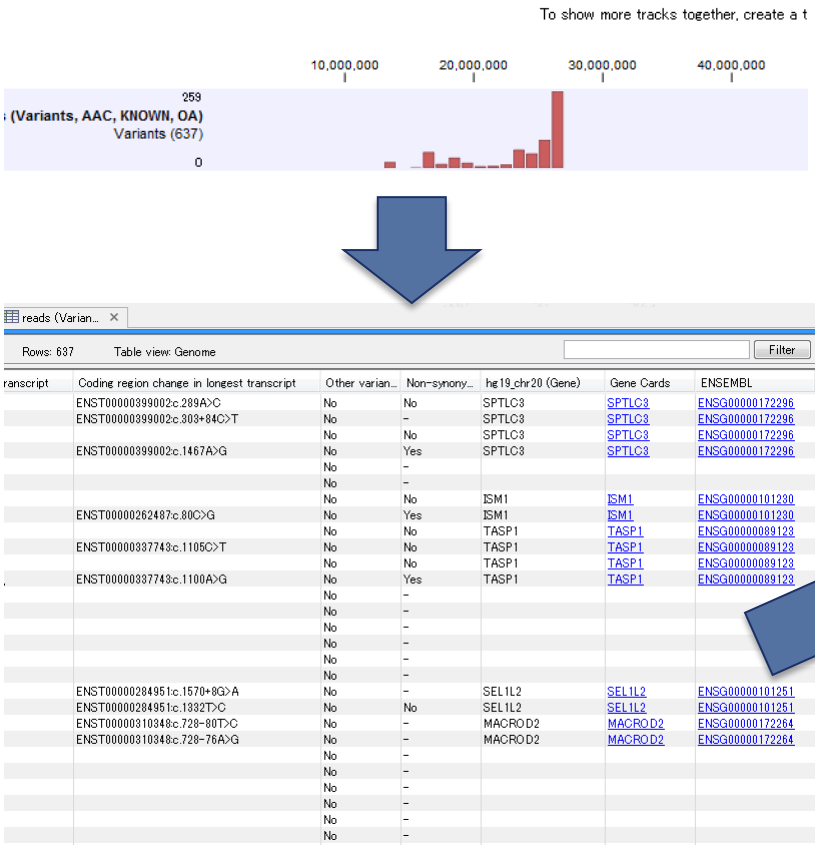
アノテーション付け：Annotate with Overlap Information

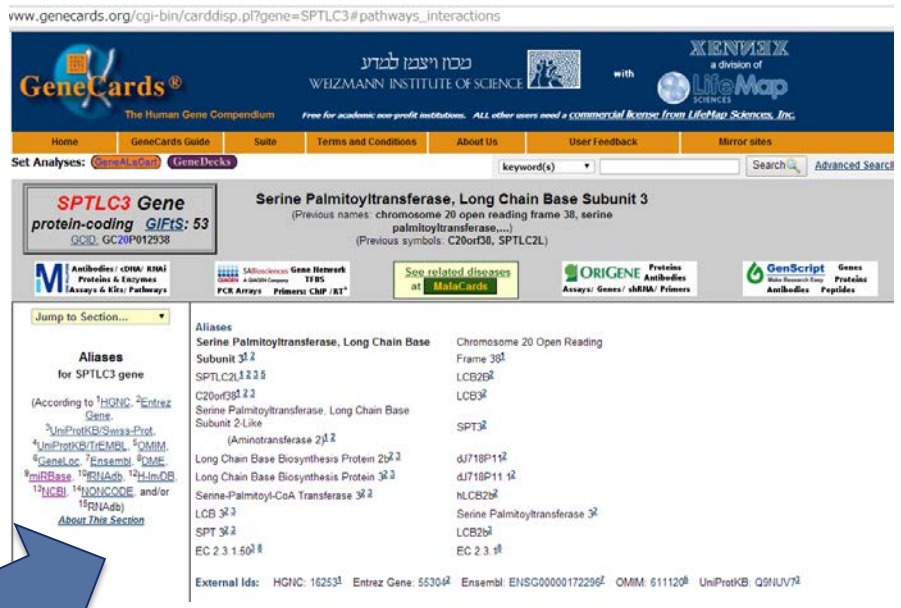


- Gene の track を選択。

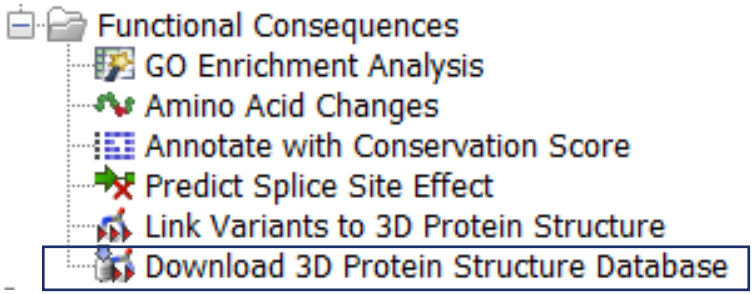
アノテーション付け：Annotate with Overlap Information

reads trimmed (Variants, AAC, KNOWN, OA)



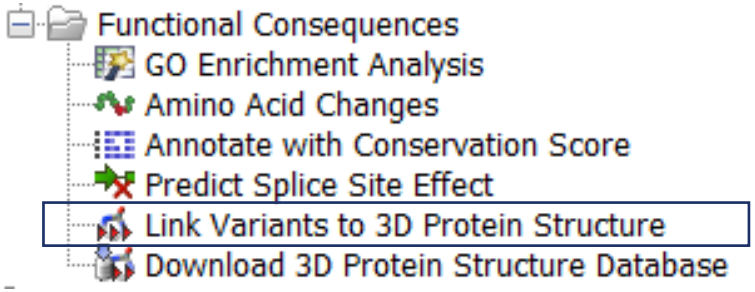


３D構造解析



Download 3D Protin Structure Database:

Resequencing Analysis > Functional Consequencesの中に含まれるこのツールでは、PDBに登録されているアミノ酸配列のデータベースをダウンロードします。この後のLink Variant to 3D Protein Structure で必要となります。



Link Variants to 3D Protein Structure:

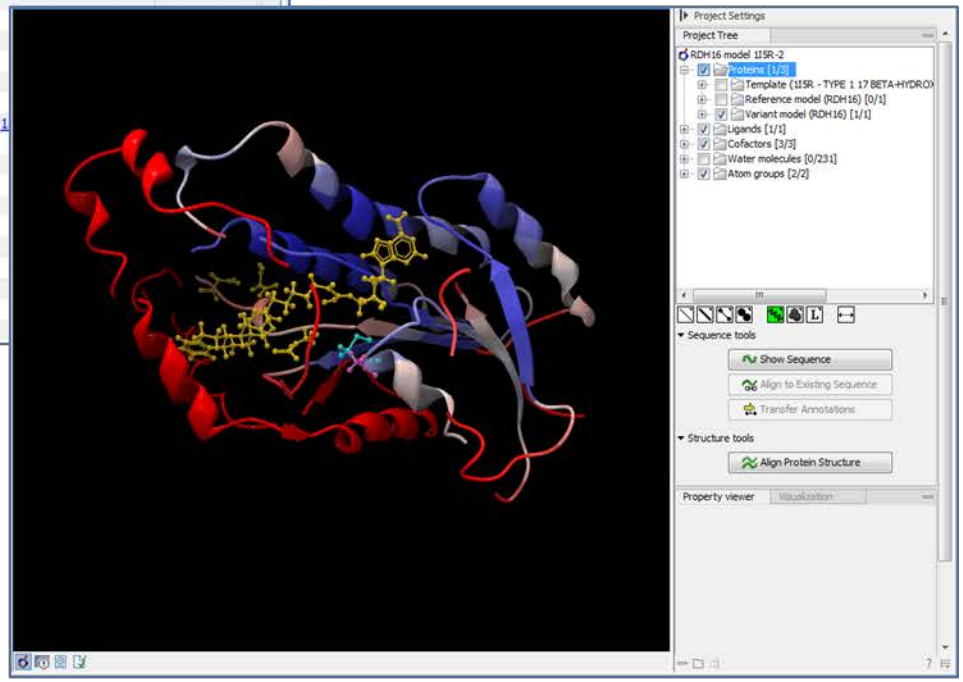
ダウンロードしたアミノ酸配列に対して、変異の検出からアミノ酸置換させた配列を使いBLAST検索をかけます。

3D構造解析：Link variants to 3D Protein Structure

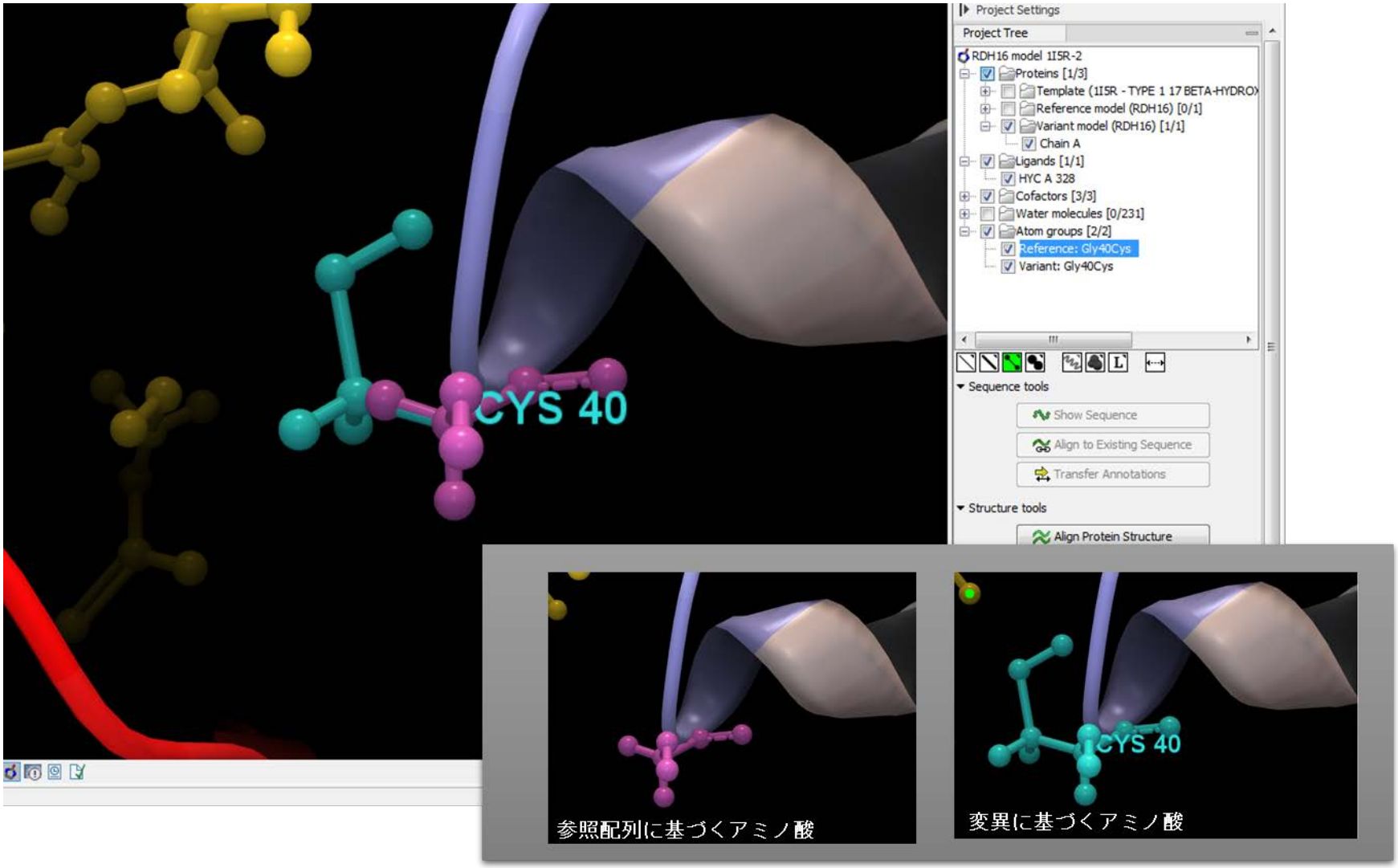
Variants (paired) Annotated Somatic Variants (LTS)

Rows: 13,182 Table view: Genome Filter

Tumour origin Cosmic_v67	NSN Clinvar...	conservation score	Link to 3D Protein Structure	dbSNP	COSMIC	PUBMED
		0.00	AGRN (synonymous)			
		0.93	C1orf159 (synonymous)			
		1.00E-3	SCNN1D (no PDB hits)			
		0.00	Outside CDS regions			
		0.02	DVL1 (no PDB hits)			
		0.83	ATAD3C Val235Met			
		0.00	ATAD3B (no PDB hits)			
		0.00	ATAD3B (no PDB hits)			
		1.00E-3	Outside CDS regions			
		0.02	ATAD3A Ser391Gly			
		0.08	Outside CDS regions			
		1.00	SSU72 (synonymous)			
		0.99	SLC35E2 (no PDB hits)			
primary, primary			NADK (no PDB hits)			
		1.00	GNB1 Asp154Glu			
		0.02	Outside CDS regions			
		2.00E-3	Outside CDS regions			
		1.00	Outside CDS regions			
		1.00	PLCH2 (no PDB hits)			
		0.00	MMEL1 Met518Thr			
		0.78	ACTRT2 (synonymous)			
		0.97	ARHGEF16 Ser360Arg			
		0.00	LRRC47 (synonymous)			
		0.95	CEP104 (nonsense)			



3D構造解析：Link variants to 3D Protein Structure



お疲れ様でした。