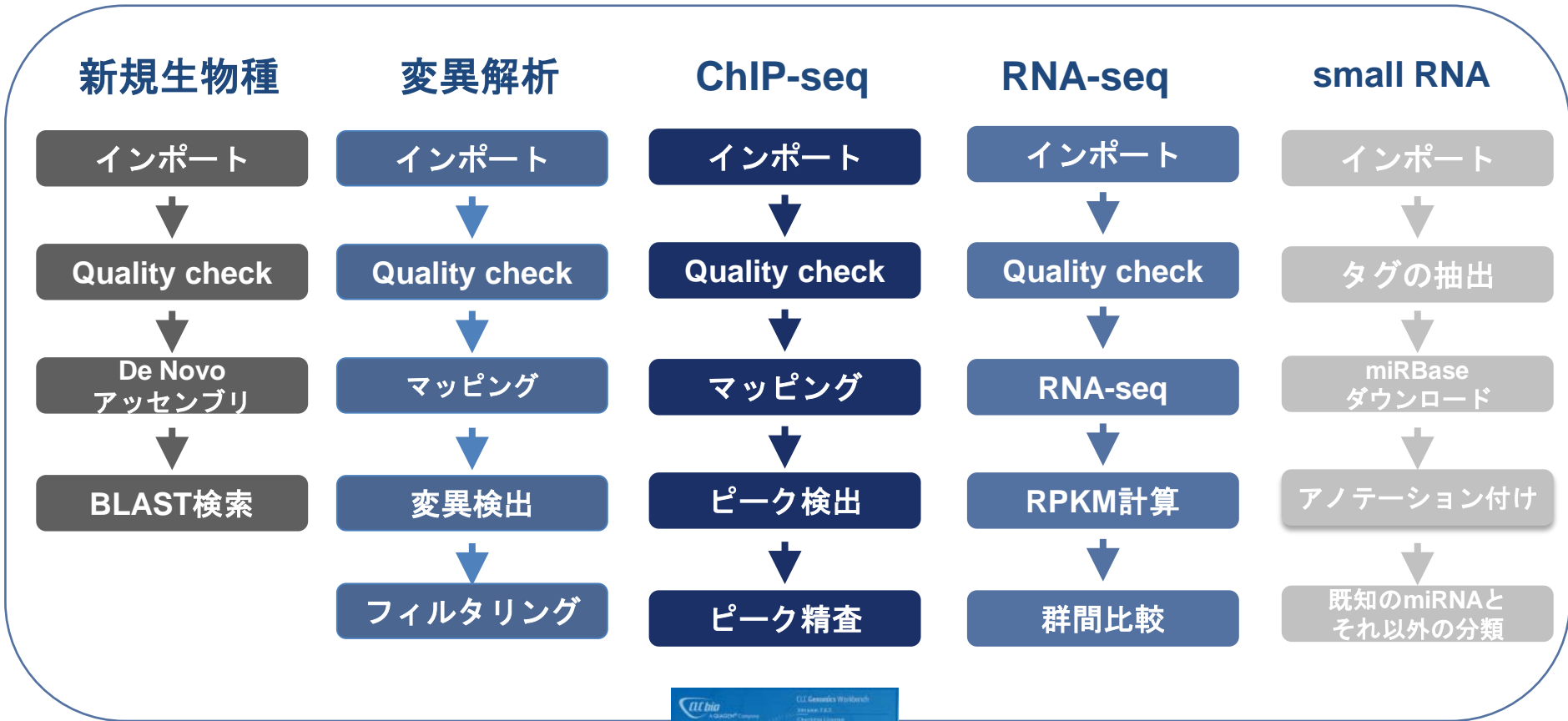


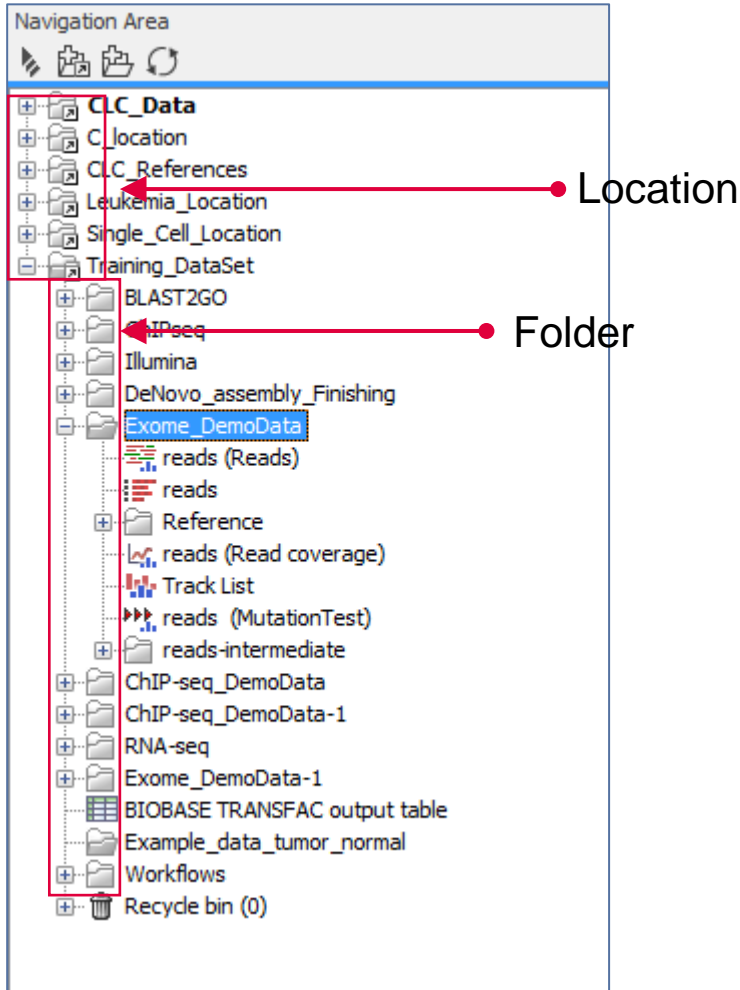


CLC Genomics Workbench ハンズオントレーニング 変異解析編

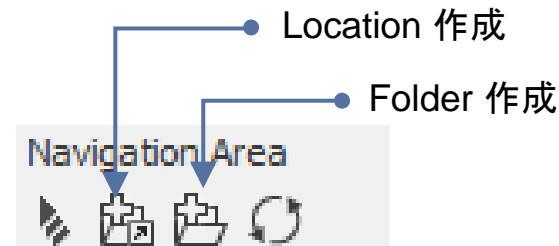
株式会社キアゲン
グローバルイフォマティクス ソリューション&サポート
アプライドアドバンストゲノミクス



データロケーション



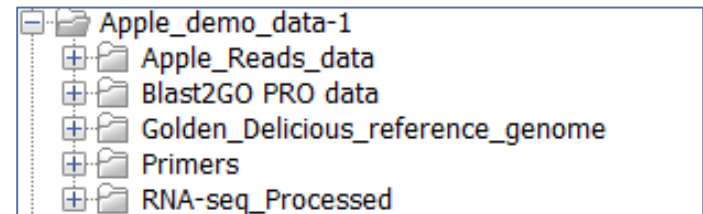
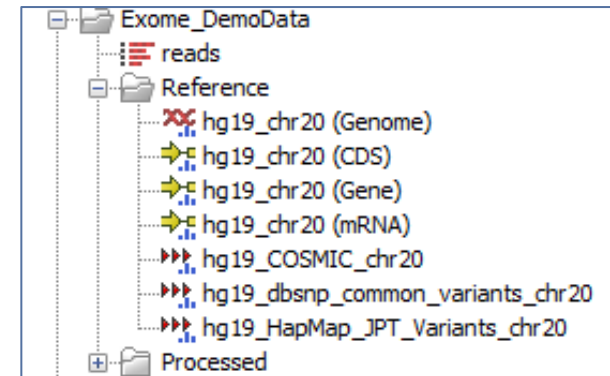
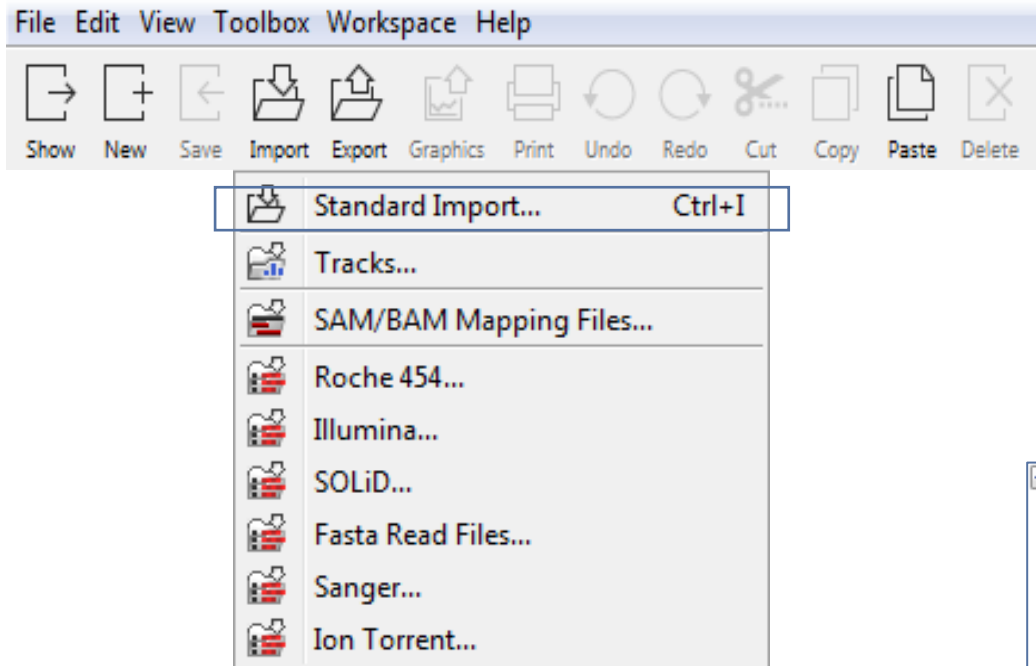
- Genomics Workbench ではデータ保存の階層のトップをLocationと呼びます。
- デフォルトのLocationはCLC_Dataが作成されていますが、左の図のようにLocationは追加可能です。
- Location の新規追加は、Navigation Area 左上のアイコンから作成可能です。シーケンスデータはサイズが大きいいため、容量が大きいディスクへLocationを作成することをお勧めします。



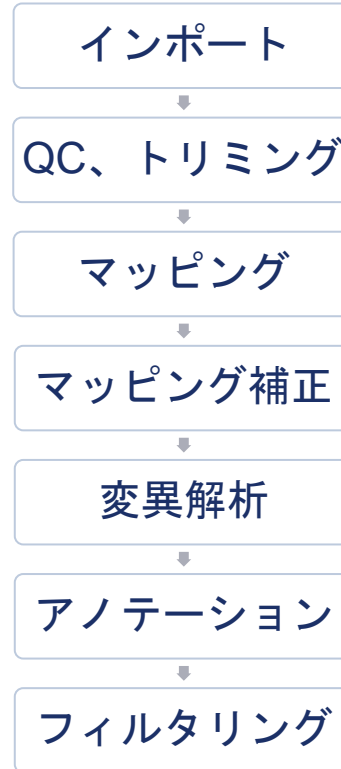
- また解析が一通り終了し、バックアップや外付けのディスクへ移動する場合は、このLocation単位での移動をお願いします。

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。



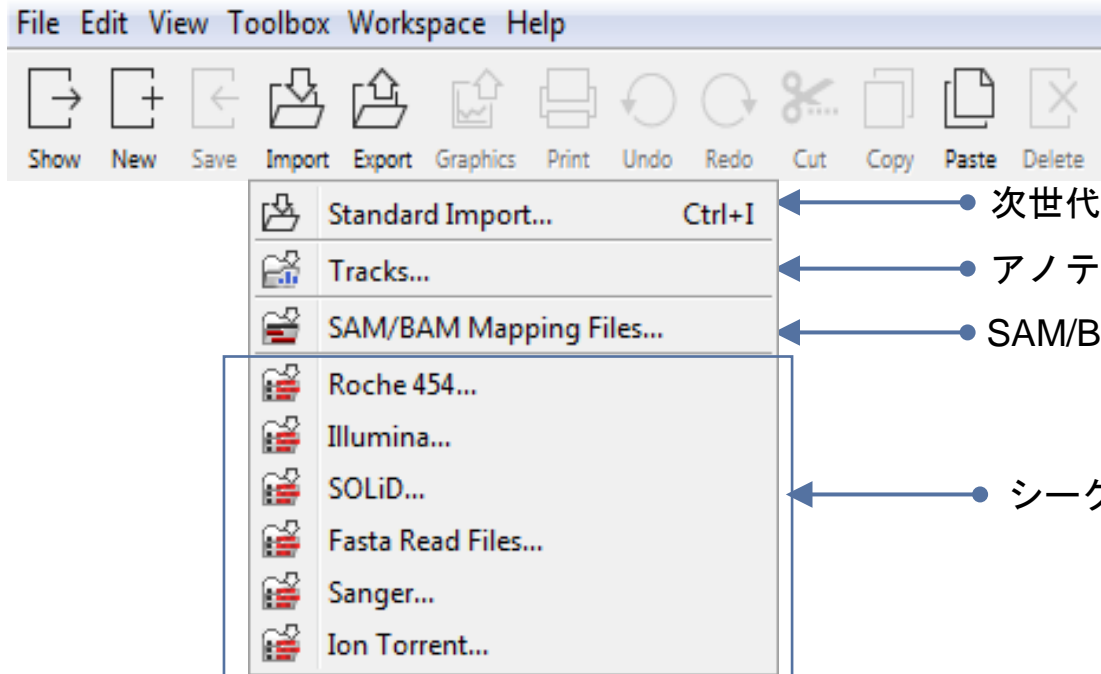
全体の流れ



CLC Genomics Workbench

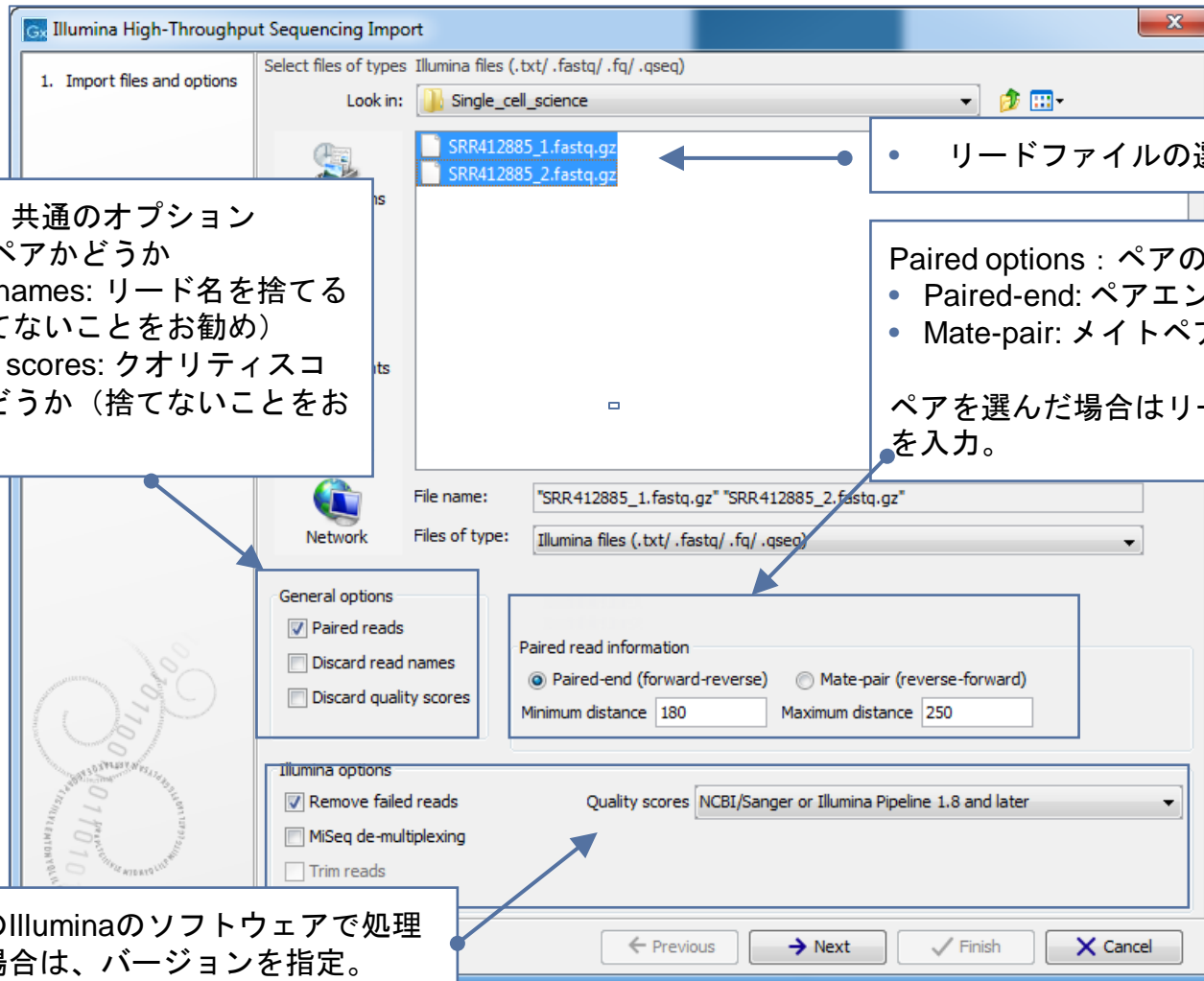
データインポート

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



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)

リードファイルの選択

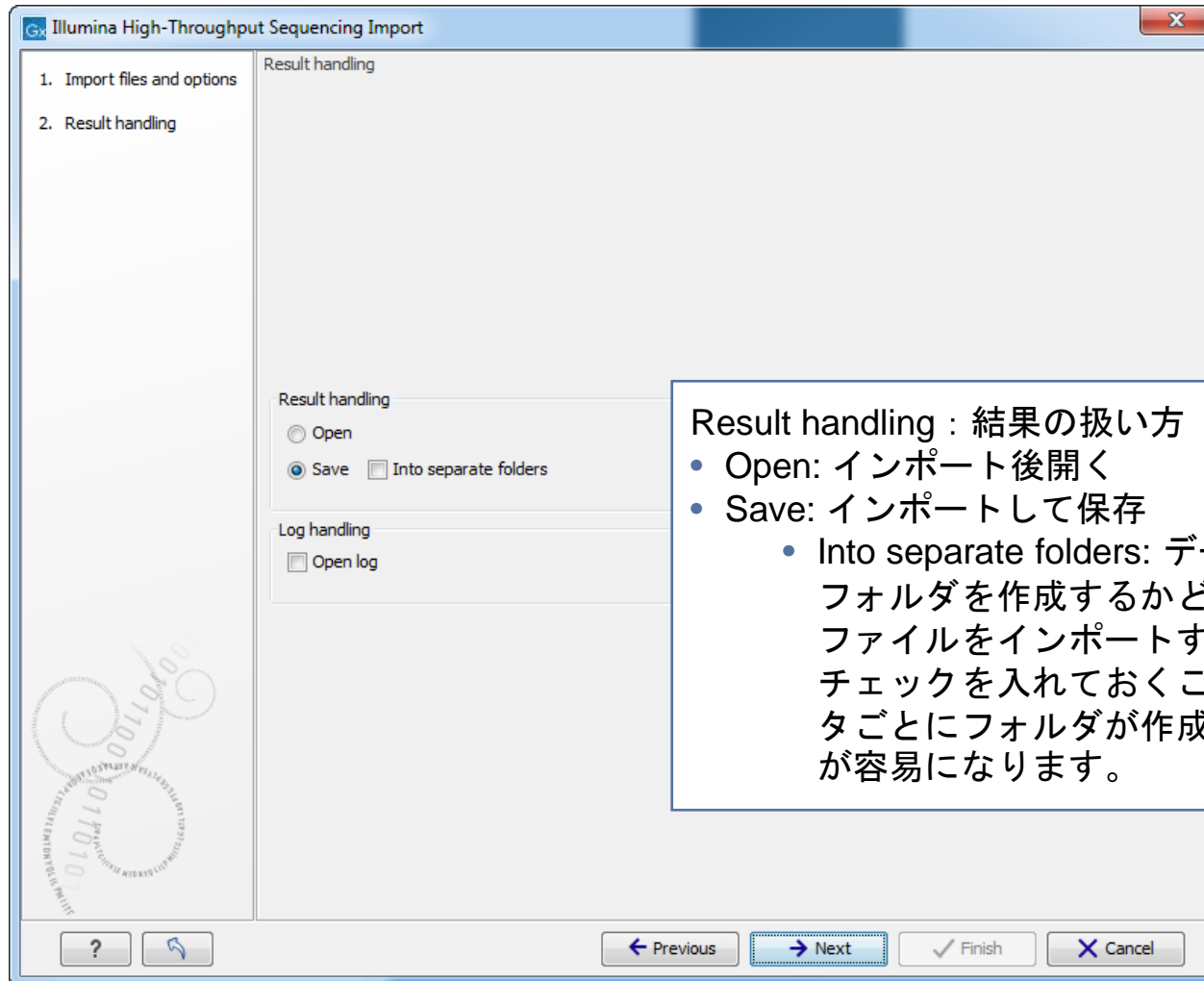
Paired options : ペアのオプション

- 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。

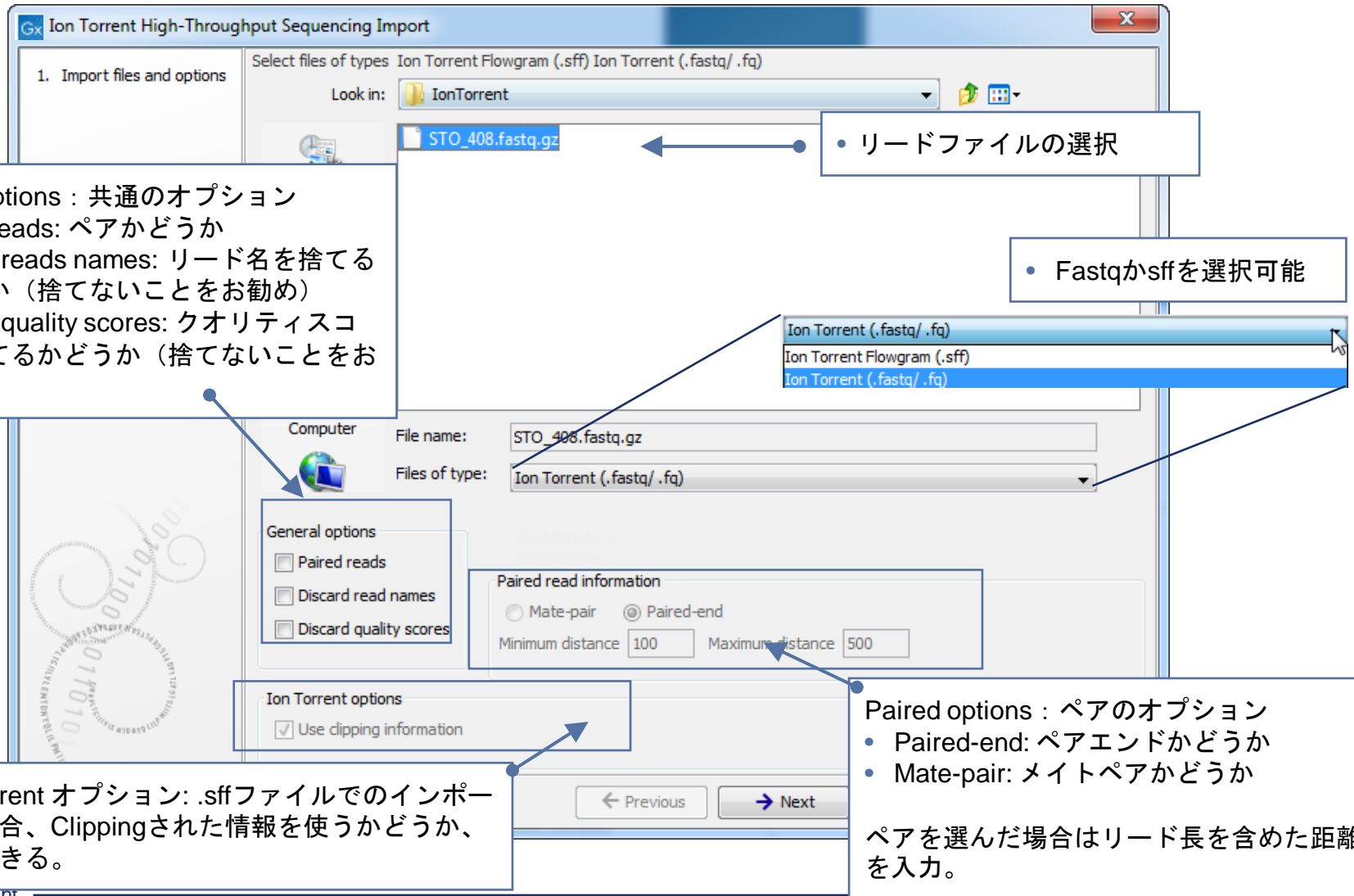
リードデータインポート：イルミナ



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



1. Import files and options

Select files of types: Ion Torrent Flowgram (.sff) Ion Torrent (.fastq/.fq)

Look in: IonTorrent

STO_408.fastq.gz

• 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)

Computer

File name: STO_408.fastq.gz

Files of type: Ion Torrent (.fastq/.fq)

General options

- Paired reads
- Discard read names
- Discard quality scores

Paired read information

Mate-pair Paired-end

Minimum distance: 100 Maximum distance: 500

Ion Torrent options

- Use clipping information

← Previous → Next

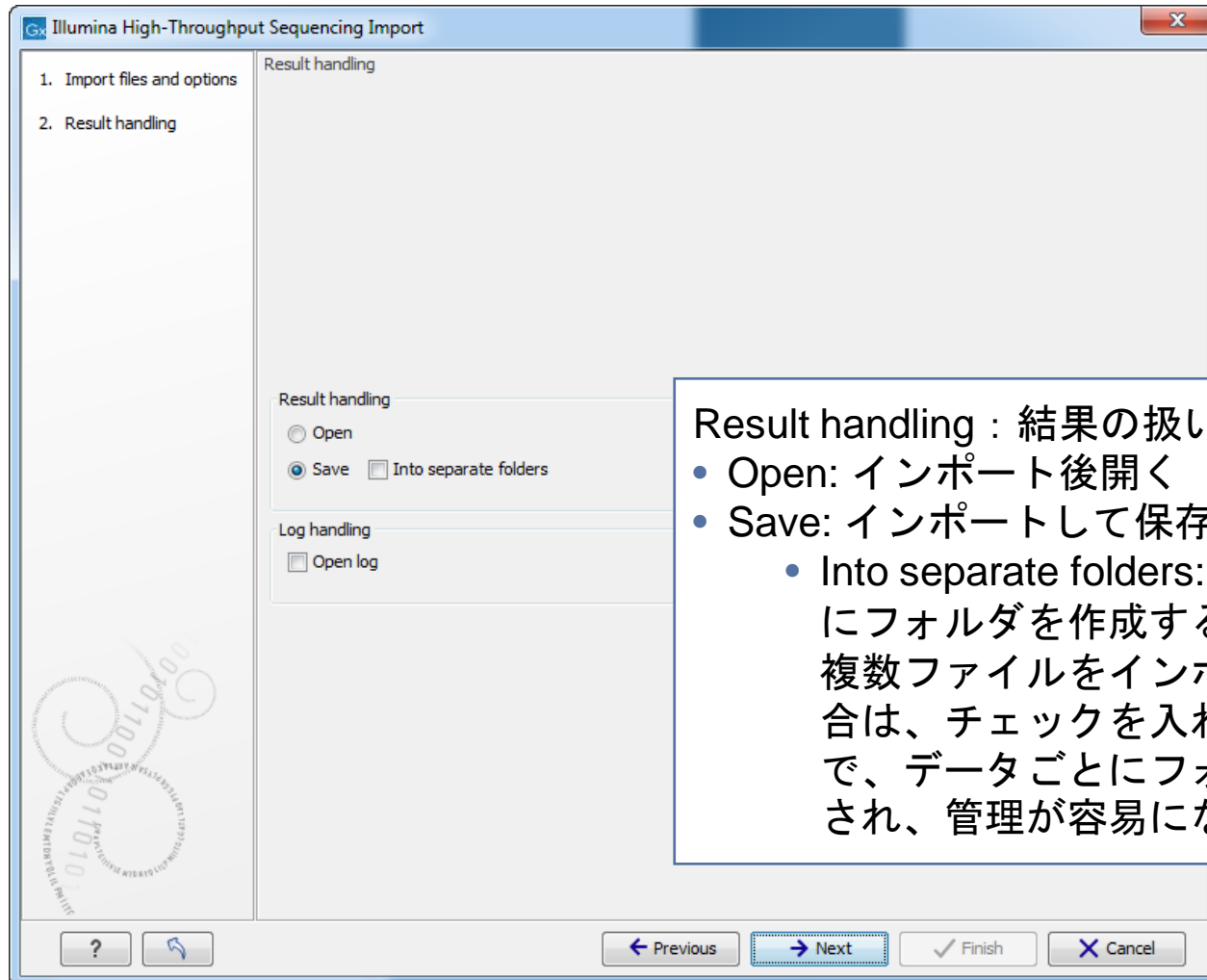
Paired options : ペアのオプション

- 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。

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。

リードデータインポート : Ion Torrent

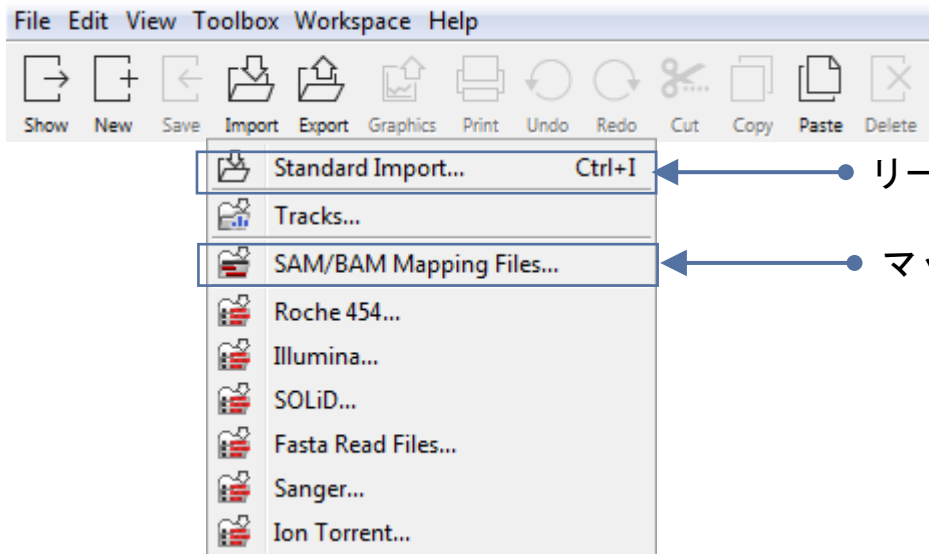


Result handling : 結果の扱い方

- 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。

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

Ion Torrentのシーケ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。

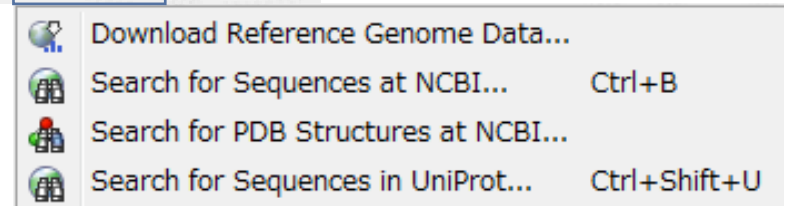
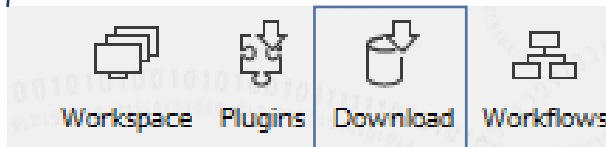
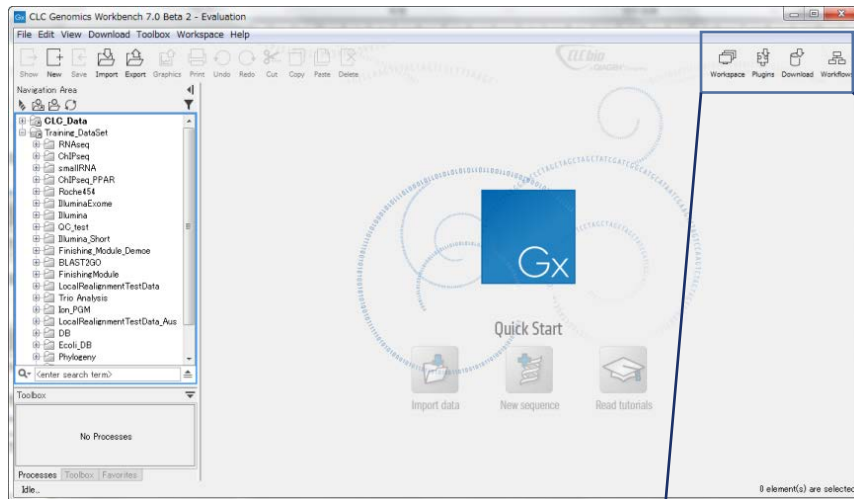


● リードデータとしてインポートされます。

● マッピングデータとしてインポートされます。

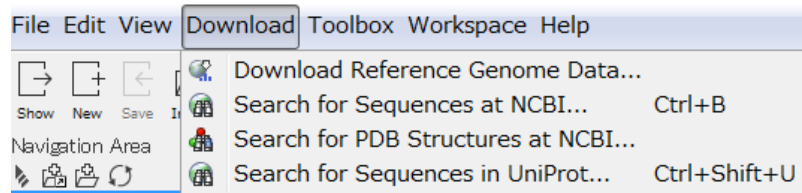
ゲノムインポート

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。

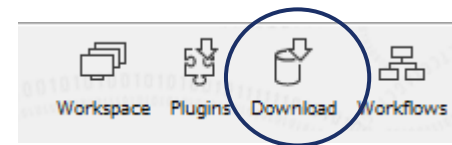


ゲノム配列の入手方法

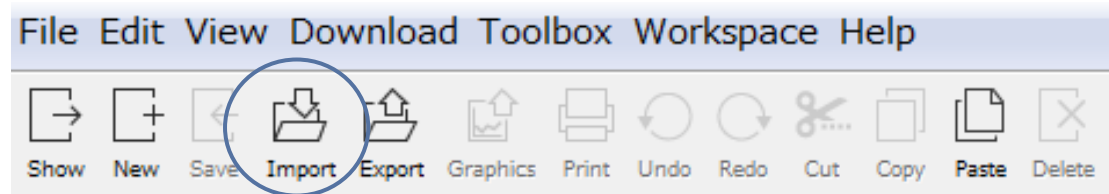
- Download 機能を用いる



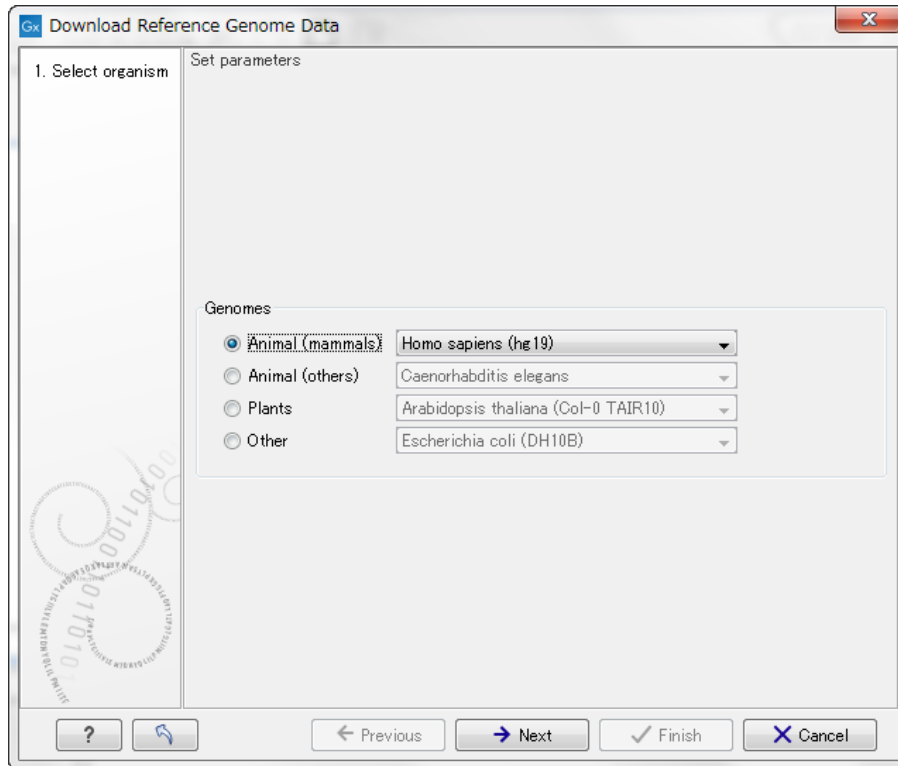
または



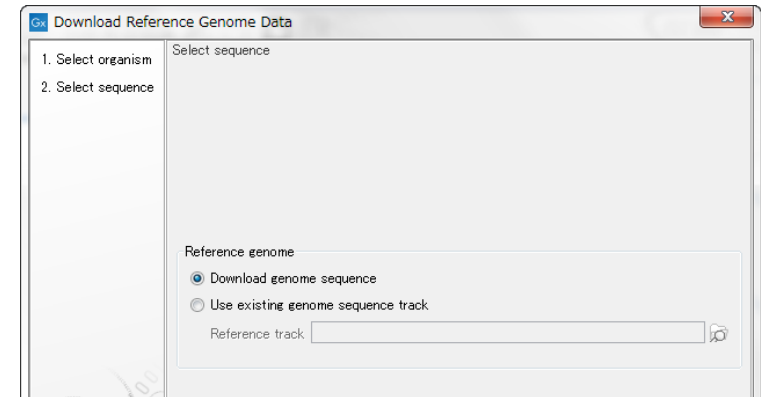
- Download サイトからダウンロードしたファイルをインポートする



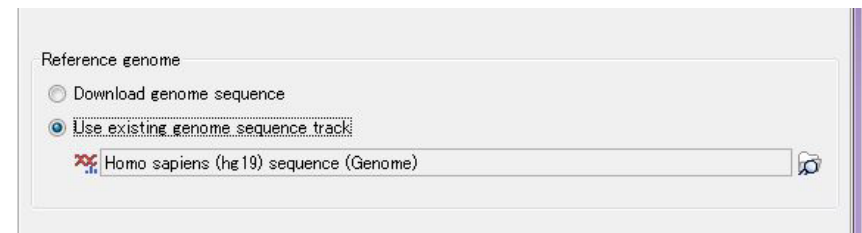
ゲノムインポート



- 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 existing 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

Download Reference Genome Data

1. Select organism
2. Select sequence
3. Select annotations

Select annotations

Download	Name	Version	Provider	Size (in Mb)
<input checked="" type="checkbox"/>	Sequence	74	Ensembl	827
<input checked="" type="checkbox"/>	Gene annotation	74	Ensembl	28
<input checked="" type="checkbox"/>	Dbsnp (common) variants	137	UCSC	528
<input checked="" type="checkbox"/>	Dbsnp variants	137	UCSC	1472
<input checked="" type="checkbox"/>	COSMIC	v67_241013	SANGER	50
<input checked="" type="checkbox"/>	Clinical variants in dbSNP		NCBI	1
<input checked="" type="checkbox"/>	HapMap Variants		Ensembl	441
<input checked="" type="checkbox"/>	1000genomes	phase1	Ensembl	1912

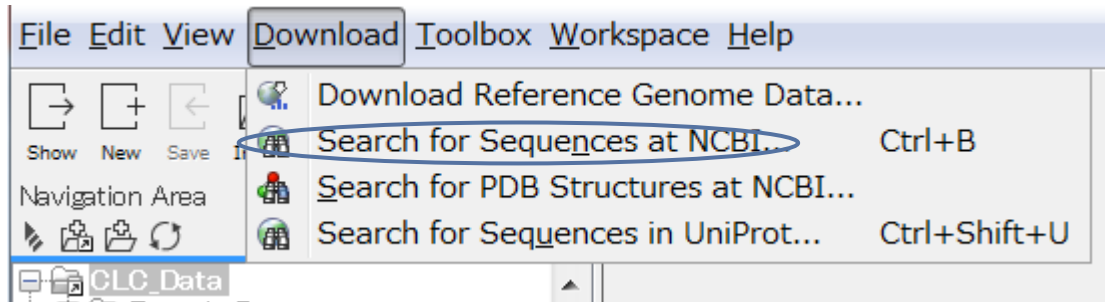
Total download size (in Mb): 5261

? ↶ ↷ ✓ ✕

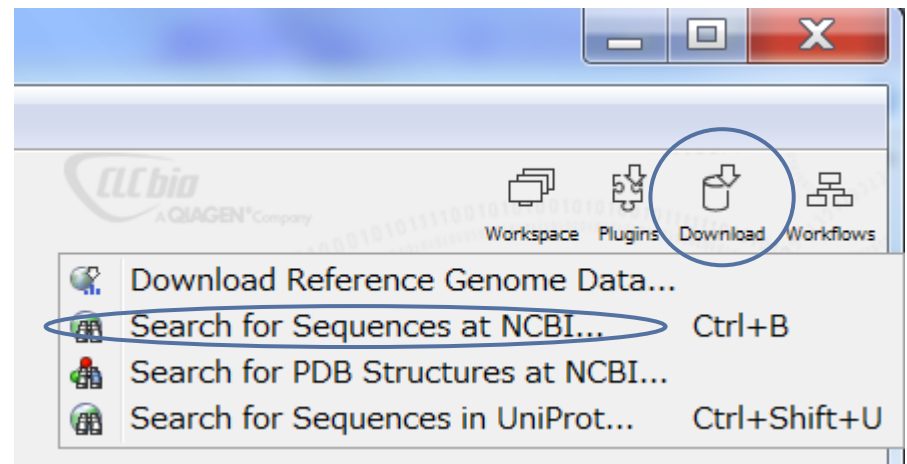
← Previous → Next ✓ Finish ✕ Cancel

- 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。

NCBIで検索してインポート

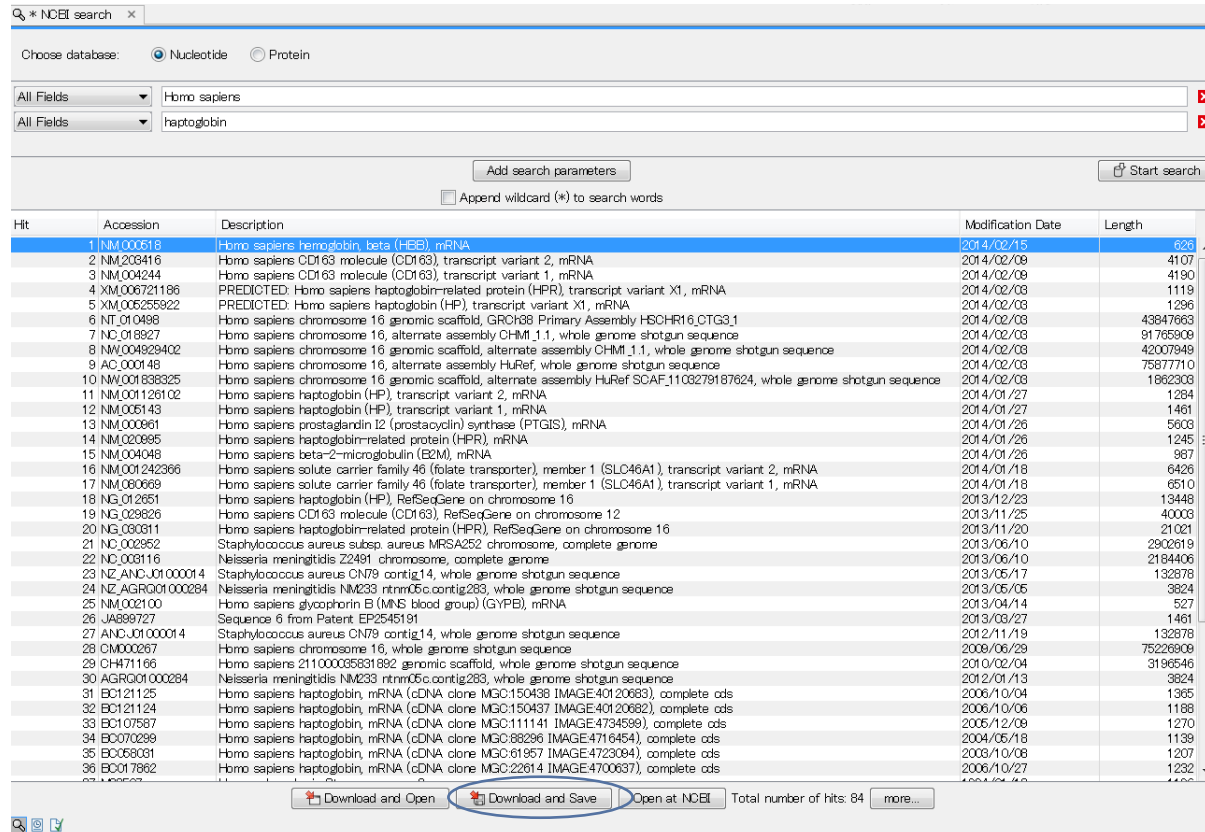


または



- 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

Search for Sequences at NCBI



NCBI search interface showing search results for 'haptoglobin' in the 'Homo sapiens' database. The search parameters are: Database: Nucleotide, Fields: All Fields, Accession: haptoglobin. The search results table is as follows:

Hit	Accession	Description	Modification Date	Length
1	NM_000518	Homo sapiens haptoglobin, beta (HEB), mRNA	2014/02/15	626
2	NM_200416	Homo sapiens CD163 molecule (CD163), transcript variant 2, mRNA	2014/02/09	4107
3	NM_004244	Homo sapiens CD163 molecule (CD163), transcript variant 1, mRNA	2014/02/09	4190
4	XM_006721186	PREDICTED: Homo sapiens haptoglobin-related protein (HPR), transcript variant X1, mRNA	2014/02/03	1119
5	NM_006255922	PREDICTED: Homo sapiens haptoglobin (HP), transcript variant X1, mRNA	2014/02/03	1296
6	NT_010488	Homo sapiens chromosome 16 genomic scaffold, GRCh38 Primary Assembly HSCHR16_CTG3_1	2014/02/03	43847663
7	NC_018927	Homo sapiens chromosome 16, alternate assembly CHM1.1.1, whole genome shotgun sequence	2014/02/03	81765908
8	NM_004829402	Homo sapiens chromosome 16 genomic scaffold, alternate assembly CHM1.1.1, whole genome shotgun sequence	2014/02/03	42007949
9	AC_000148	Homo sapiens chromosome 16, alternate assembly HuRef, whole genome shotgun sequence	2014/02/03	75877710
10	NM_001838325	Homo sapiens chromosome 16 genomic scaffold, alternate assembly HuRef SCAF_1109279187624, whole genome shotgun sequence	2014/02/03	1862308
11	NM_001126102	Homo sapiens haptoglobin (HP), transcript variant 2, mRNA	2014/01/27	1284
12	NM_006143	Homo sapiens haptoglobin (HP), transcript variant 1, mRNA	2014/01/27	1461
13	NM_000981	Homo sapiens prostaglandin I2 (prostaglandin) synthase (PTGIS), mRNA	2014/01/26	5903
14	NM_020895	Homo sapiens haptoglobin-related protein (HPR), mRNA	2014/01/26	1245
15	NM_004048	Homo sapiens beta-2-microglobulin (E2M), mRNA	2014/01/26	987
16	NM_001242366	Homo sapiens solute carrier family 46 (folate transporter), member 1 (SLC46A1), transcript variant 2, mRNA	2014/01/18	6426
17	NM_080669	Homo sapiens solute carrier family 46 (folate transporter), member 1 (SLC46A1), transcript variant 1, mRNA	2014/01/18	6510
18	NG_012651	Homo sapiens haptoglobin (HP), RefSeqGene on chromosome 16	2013/12/23	13448
19	NG_029826	Homo sapiens CD163 molecule (CD163), RefSeqGene on chromosome 12	2013/11/25	40003
20	NG_030311	Homo sapiens haptoglobin-related protein (HPR), RefSeqGene on chromosome 16	2013/11/20	21021
21	NC_002852	Staphylococcus aureus subsp. aureus MRSA252 chromosome, complete genome	2013/06/10	2902619
22	NC_003116	Neisseria meningitidis Z2491 chromosome, complete genome	2013/06/10	2184406
23	NZ_LNCJ01000014	Staphylococcus aureus CN79 contig14, whole genome shotgun sequence	2013/05/17	132678
24	NZ_AGRO01000284	Neisseria meningitidis NM233 ntmm5c.contig283, whole genome shotgun sequence	2013/05/06	3824
25	NM_002100	Homo sapiens glycoporphin B (MNS blood group) (GYPB), mRNA	2013/04/14	527
26	J4889727	Sequence 6 from Patent EP2545191	2013/03/27	1461
27	ANC_01000014	Staphylococcus aureus CN79 contig14, whole genome shotgun sequence	2012/11/19	132678
28	CM000297	Homo sapiens chromosome 16, whole genome shotgun sequence	2008/06/29	75298909
29	CH471166	Homo sapiens 21100005831892 genomic scaffold, whole genome shotgun sequence	2010/02/04	3196546
30	AGRO01000284	Neisseria meningitidis NM233 ntmm5c.contig283, whole genome shotgun sequence	2012/01/13	3824
31	BC121125	Homo sapiens haptoglobin, mRNA (cDNA clone MGC:150438 IMAGE:40120683), complete cds	2006/10/04	1365
32	BC121124	Homo sapiens haptoglobin, mRNA (cDNA clone MGC:150437 IMAGE:40120682), complete cds	2006/10/06	1188
33	BC107587	Homo sapiens haptoglobin, mRNA (cDNA clone MGC:111141 IMAGE:4734589), complete cds	2005/12/09	1270
34	BC070289	Homo sapiens haptoglobin, mRNA (cDNA clone MGC:88296 IMAGE:4716454), complete cds	2004/05/18	1139
35	BC058031	Homo sapiens haptoglobin, mRNA (cDNA clone MGC:61957 IMAGE:4723084), complete cds	2003/10/08	1207
36	BC017882	Homo sapiens haptoglobin, mRNA (cDNA clone MGC:22614 IMAGE:4700637), complete cds	2006/10/27	1232

- 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

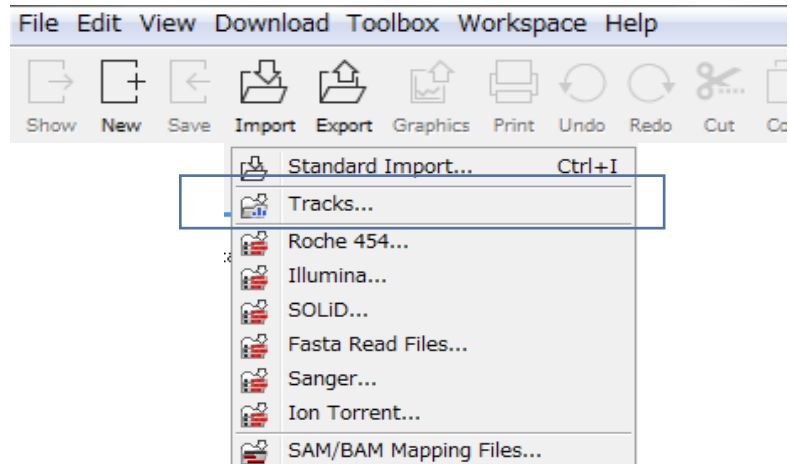
アノテーションインポート

- 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 dump
 - COSMIC variation database

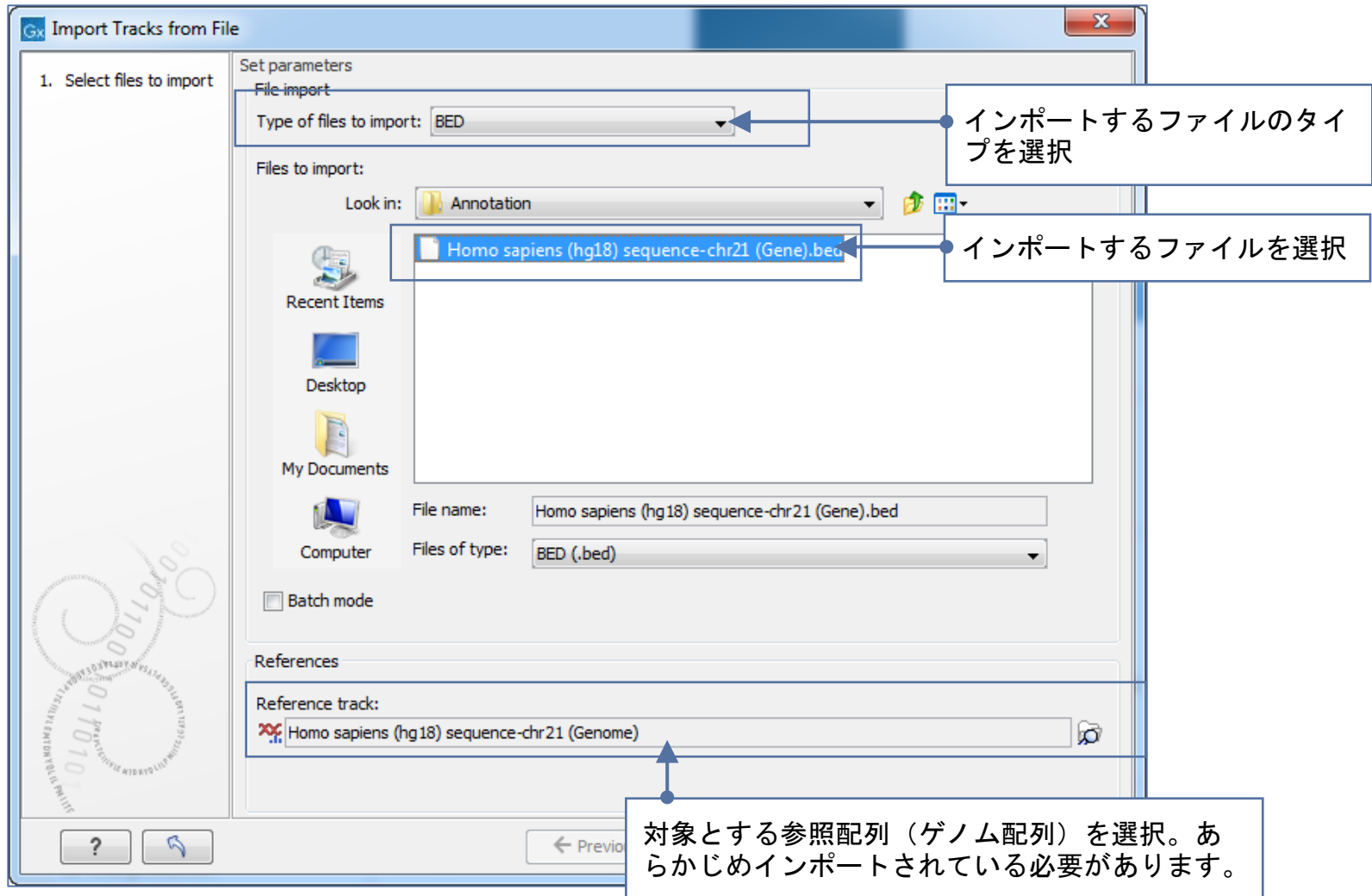
※変異のデータについても、アノテーションとして自分の変異へアノテーションとして情報の追加や比較ができるため、アノテーションのインポート可能フォーマットに含めています。

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



The screenshot shows the 'Import Tracks from File' dialog box. It is divided into several sections:

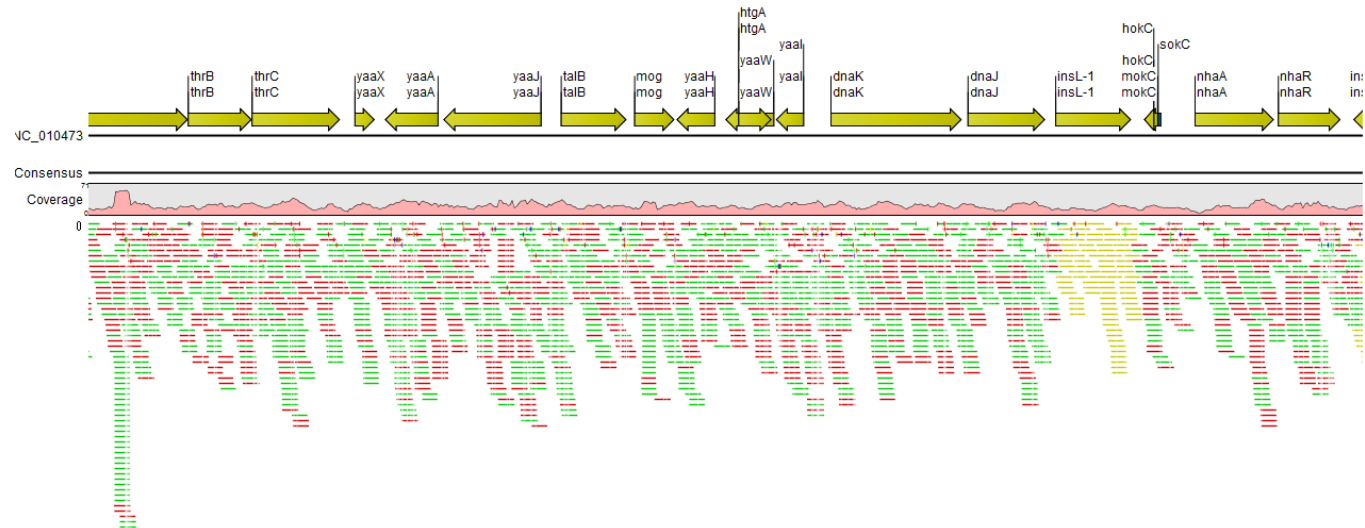
- 1. Select files to import:** A sidebar on the left with icons for Recent Items, Desktop, My Documents, and Computer.
- Set parameters:**
 - File import:** A dropdown menu for 'Type of files to import' is set to 'BED'. An annotation points to this dropdown with the text: 'インポートするファイルのタイプを選択' (Select the type of file to import).
 - Files to import:** A file list showing 'Homo sapiens (hg18) sequence-chr21 (Gene).bed'. An annotation points to this file with the text: 'インポートするファイルを選択' (Select the file to import).
 - Look in:** A dropdown menu is set to 'Annotation'.
 - File name:** A text field containing 'Homo sapiens (hg18) sequence-chr21 (Gene).bed'.
 - Files of type:** A dropdown menu set to 'BED (.bed)'.
 - Batch mode
- References:**
 - Reference track:** A dropdown menu showing 'Homo sapiens (hg18) sequence-chr21 (Genome)'. An annotation points to this dropdown with the text: '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。' (Select the reference sequence (genome sequence) to be targeted. It must be imported in advance).

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。

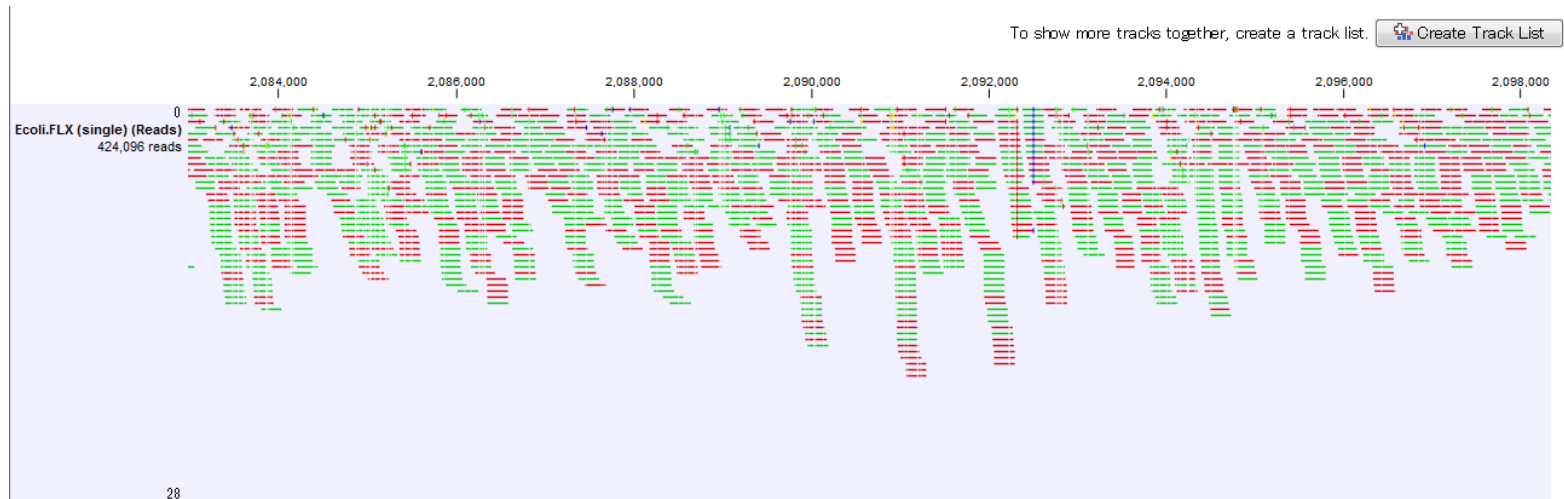
 reads mapping



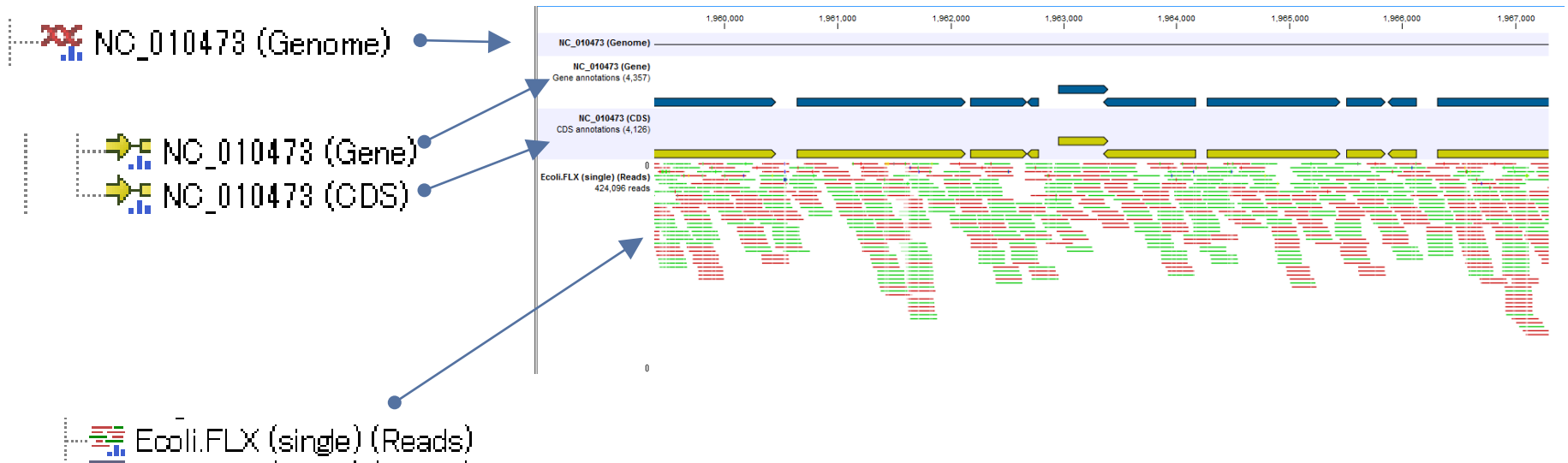
57

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。




reads (Reads)



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。











スタンドアロンフォーマット

 Homo sapiens (hg19) sequence-1	染色体のセットやリード配列など配列のセット
 chr1	染色体1本など1つの配列
 reads mapping	リードマッピング

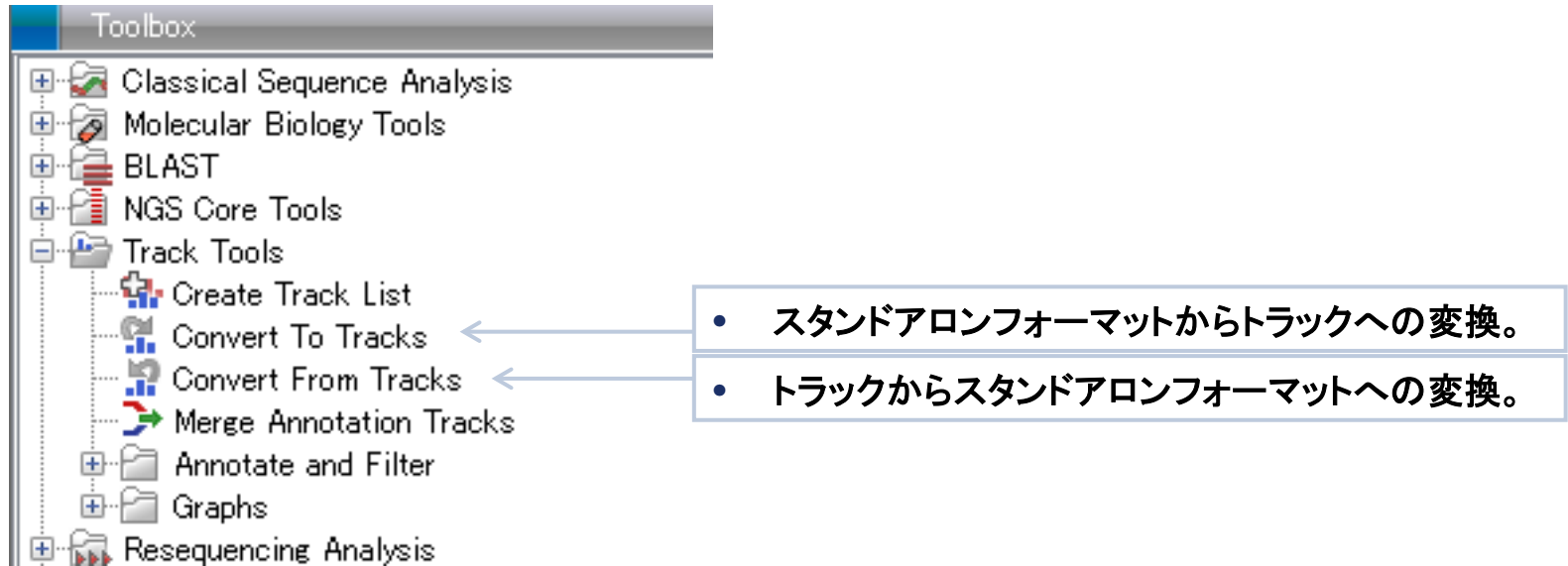
トラックフォーマット

青いヒストグラムが目印

 Homo sapiens (hg19) sequence	ゲノムTrack
 Homo sapiens (hg19)_CDS	
 Homo sapiens (hg19)_Exon	
 Homo sapiens (hg19)_Gene	アノテーションTrack
 Homo sapiens (hg19)_mRNA	
 Homo sapiens (hg19)_Transcript	
 Homo sapiens (hg19) COSMIC	変異Track
 reads (Reads)	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。

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。



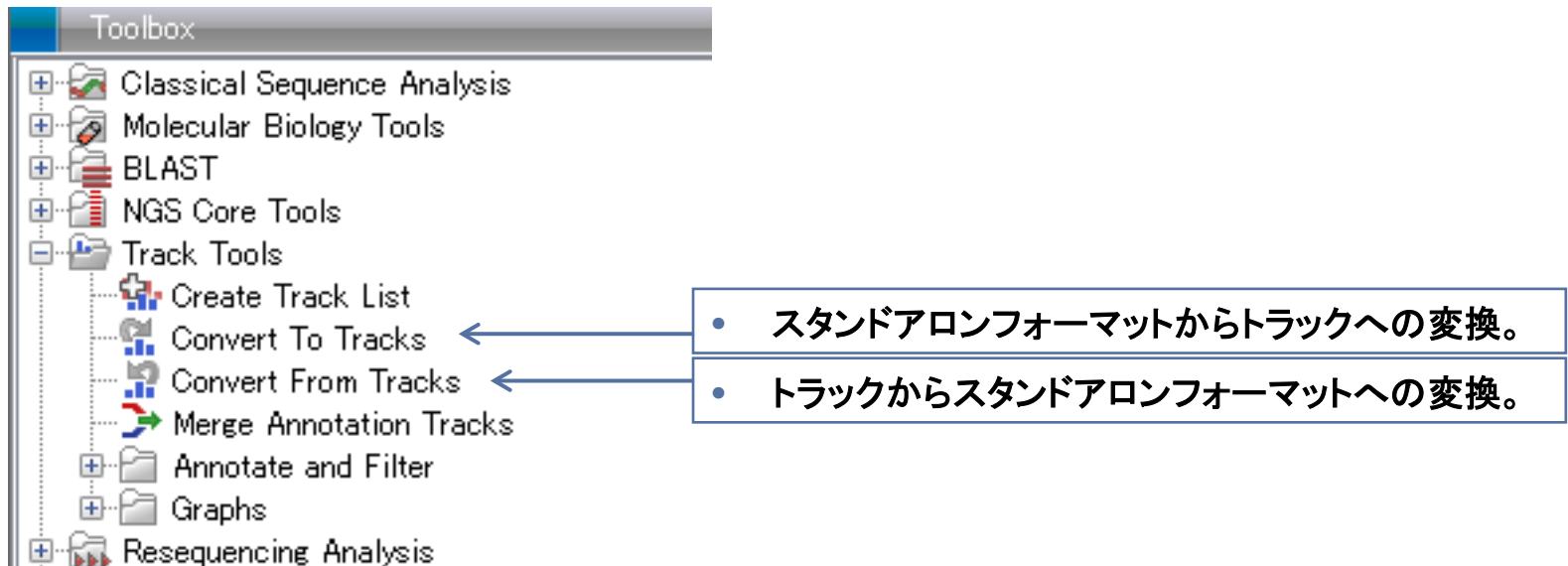
The screenshot shows the 'Toolbox' window with the following items listed:

- Classical Sequence Analysis
- Molecular Biology Tools
- BLAST
- NGS Core Tools
- Track Tools (expanded)
 - Create Track List
 - Convert To Tracks
 - Convert From Tracks
 - Merge Annotation Tracks
- Annotate and Filter
- Graphs
- Resequencing Analysis

Two callout boxes on the right provide descriptions for the highlighted tools:

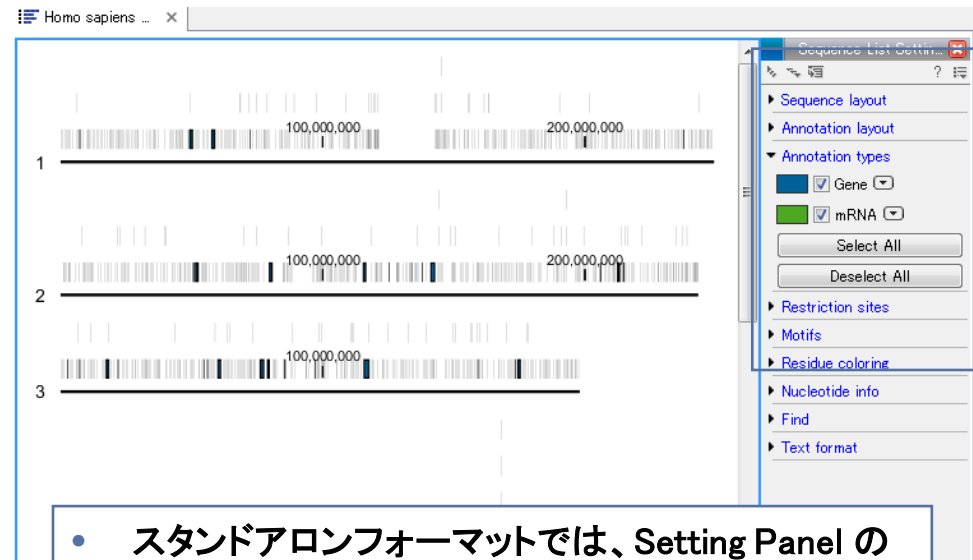
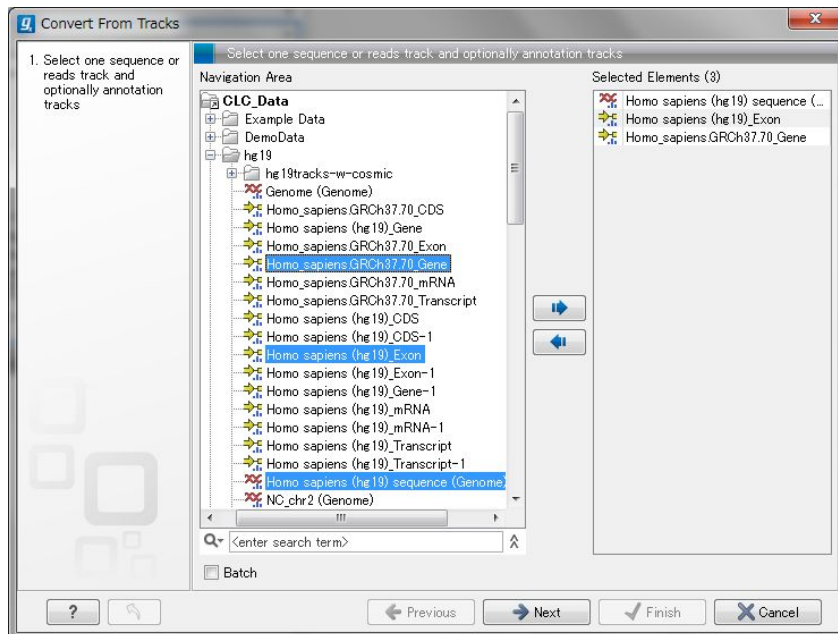
- スタンドアロンフォーマットからトラックへの変換。
- トラックからスタンドアロンフォーマットへの変換。

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

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・組み合わせでトリミング。

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。

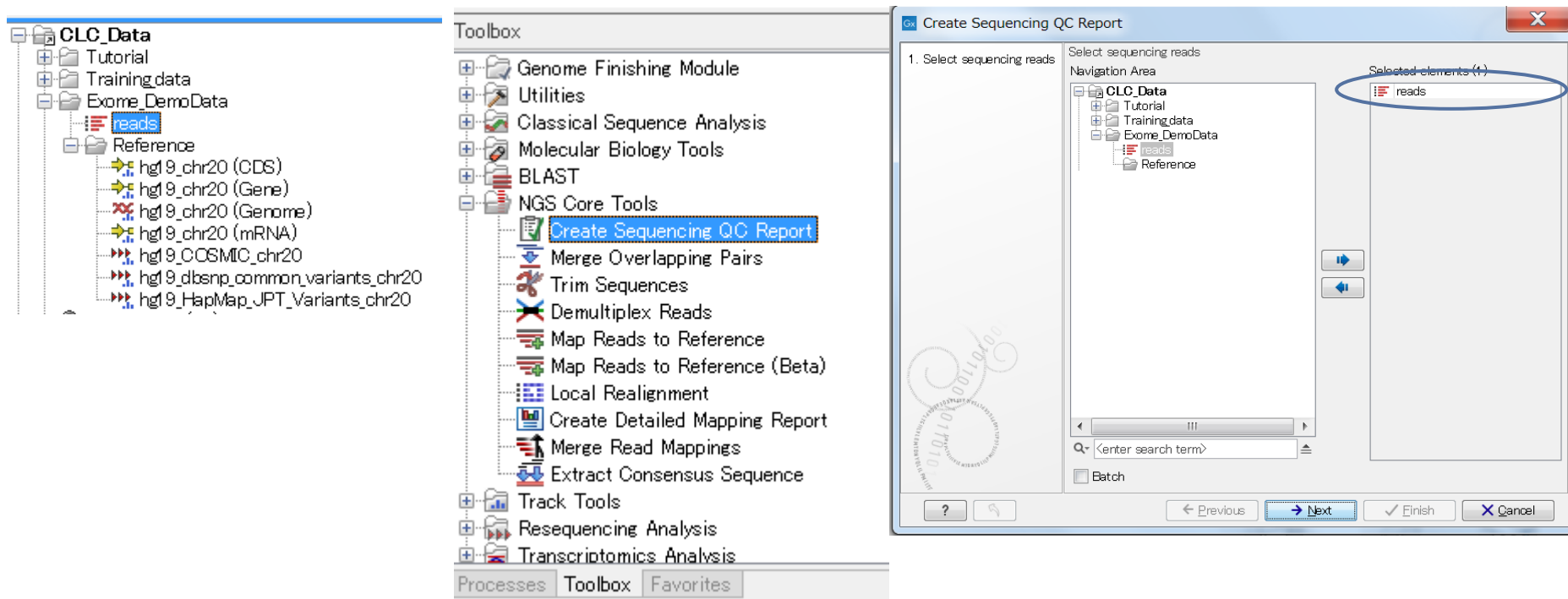
クオリティスコア

シーケ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。Phred 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。

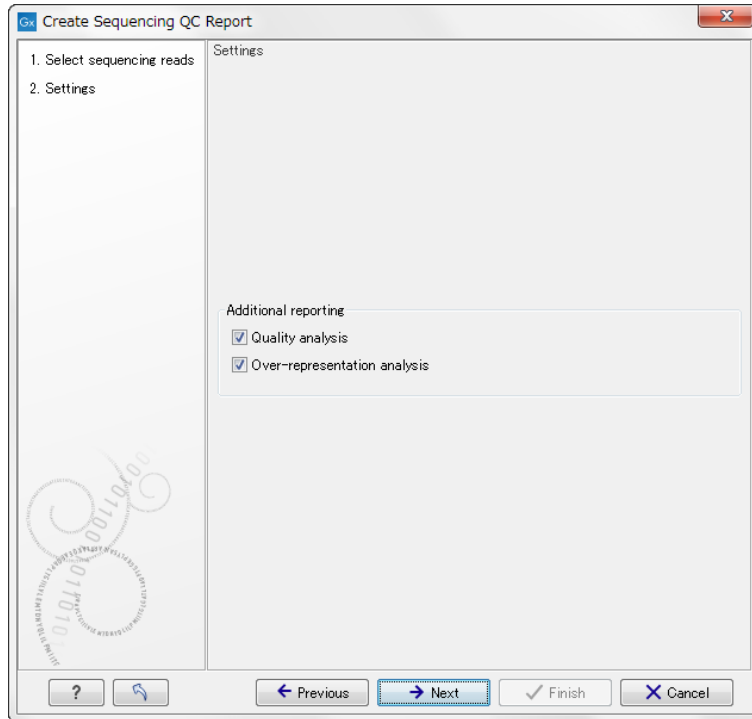
$$\text{PhredScore} = -10 \log_{10} P_{err}$$

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%

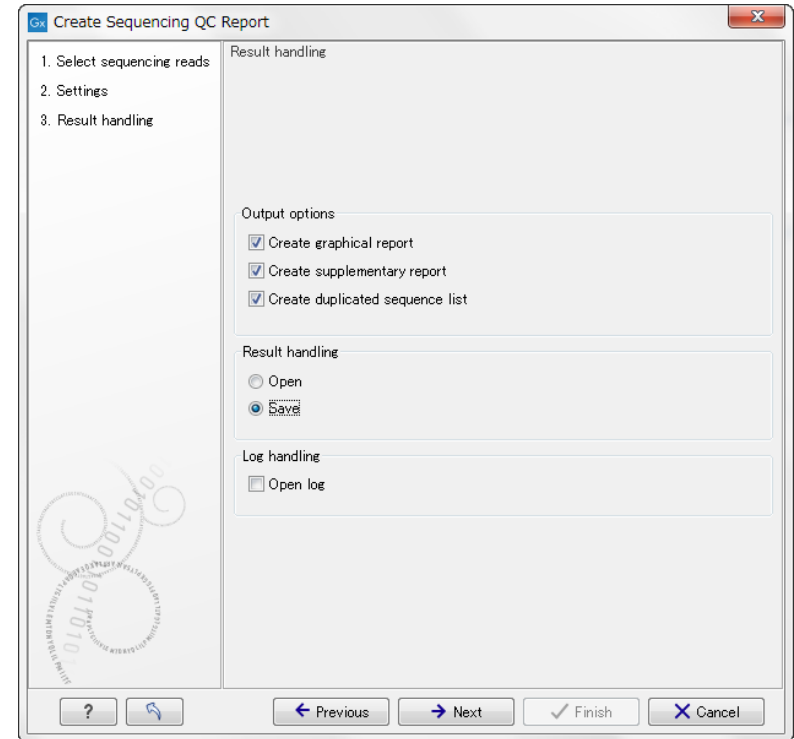


The screenshot displays the CLC software interface during the 'Create Sequencing QC Report' wizard. On the left, the project tree shows 'CLC_Data' with 'Reference' expanded and 'reads' selected. The central 'Toolbox' has 'Create Sequencing QC Report' highlighted under the 'NGS Core Tools' category. The right-hand wizard window is at step 1, 'Select sequencing reads', where the 'Navigation Area' shows the project structure and the 'Selected elements (1)' list contains 'reads'.

- 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。



- 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。

- reads - graphical QC report
- reads - supplementary QC report
- reads - duplicated sequences QC report

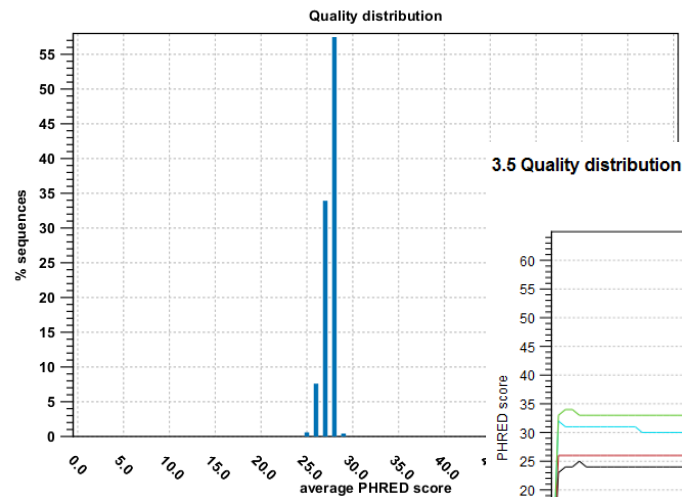
Report Settings

Table of Contents

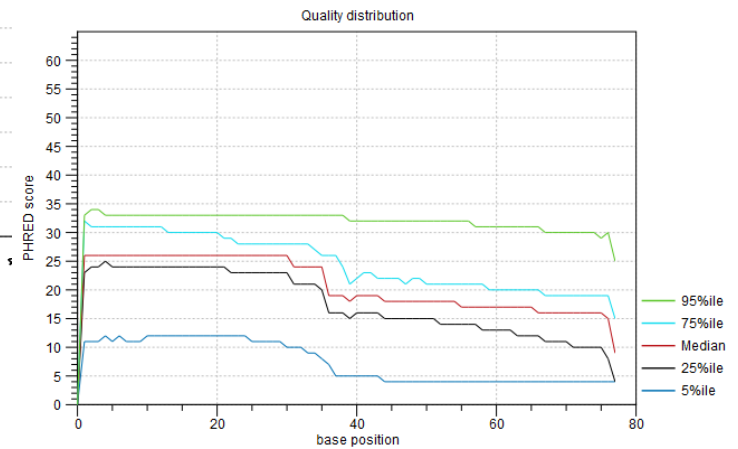
- 1 Summary
- ▼ 2 Per-sequence analysis
 - 2.1 Lengths distribution
 - 2.2 GC-content
 - 2.3 Ambiguous base-content
 - 2.4 Quality distribution
- ▼ 3 Per-base analysis
 - 3.1 Coverage
 - 3.2 Nucleotide contributions
 - 3.3 GC-content
 - 3.4 Ambiguous base-content
 - 3.5 Quality distribution
- ▼ 4 Over-representation analyses
 - 4.1 Enriched 5mers
 - 4.2 Sequence duplication levels
 - 4.3 Duplicated sequences

Text format

2.4 Quality distribution



3.5 Quality distribution



Base-quality distribution along the base positions.

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

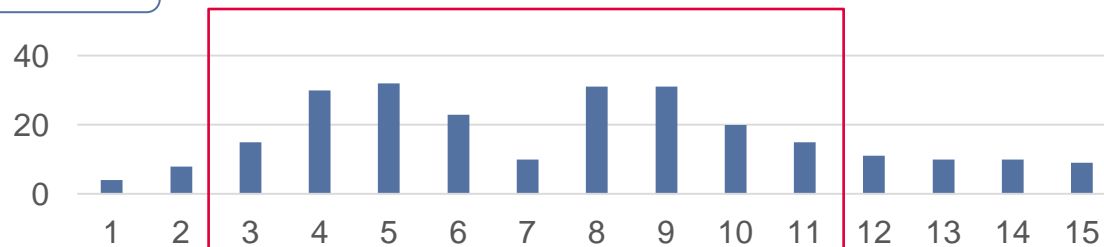
リード配列	G	C	C	C	A	T	G	T	T	C	G	A	T	G	C
Phred score	4	8	15	30	32	23	10	31	31	20	15	11	10	10	9
p値	0.40	0.16	0.03	0.00	0.00	0.01	0.10	0.00	0.00	0.01	0.03	0.08	0.10	0.10	0.13
Limit - p値 (D)	-0.35	-0.11	0.02	0.05	0.05	0.04	-0.05	0.05	0.05	0.04	0.02	-0.03	-0.05	-0.05	-0.08
(D)の累積和	0.00	0.00	0.02	0.07	0.12	0.16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

Limit = 0.05の場合

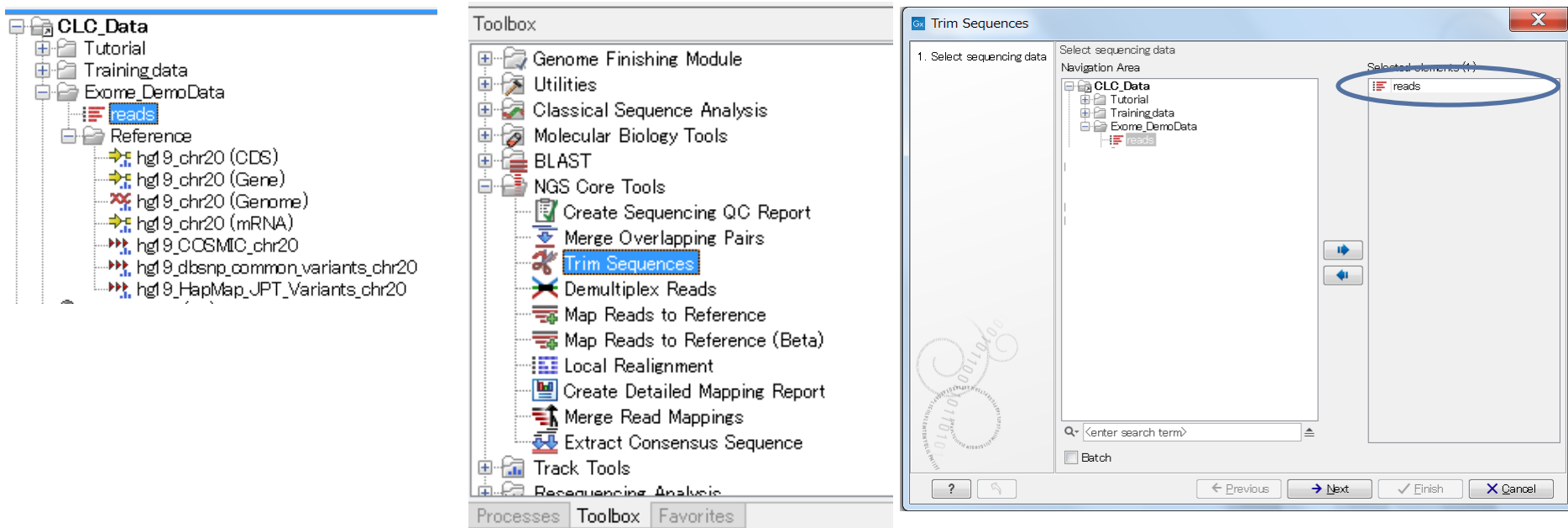
↑
 スタート点：
 累積和が0より大き
 くなった塩基

↑
 終了点：
 累積和が最大を
 示す塩基

Phred score の棒グラフ



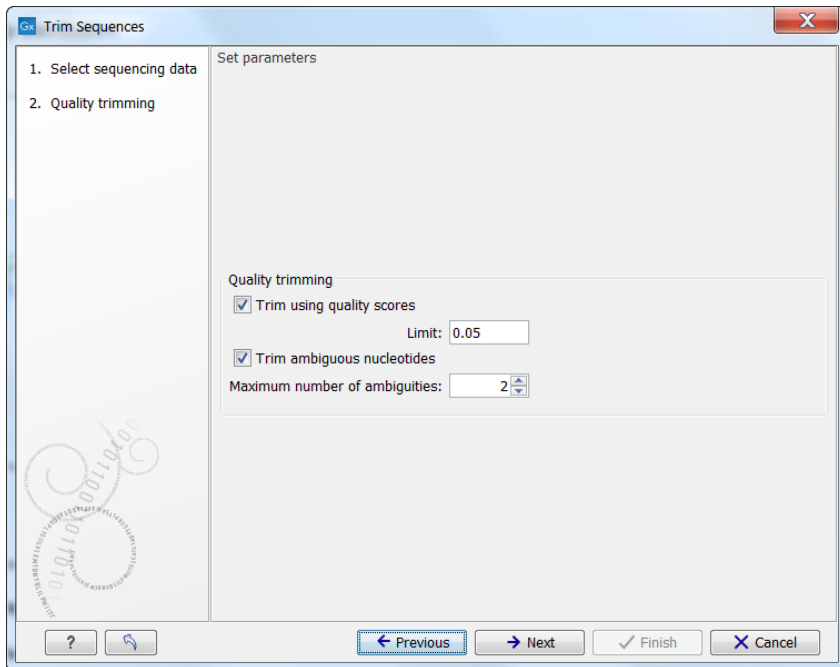
グラフより、ある程度クオリティが高くなった場所からリードを使い、クオリティが連続して悪くなっている箇所からリードをトリミングしていることがわかる。
 ※途中、1塩基のみクオリティが低いような場合は、必ずしもトリミングされない。これはできるだけリードを長く保とうとするため。



The screenshot displays the QIAGEN CLC software interface for the 'Trim Sequences' workflow. On the left, a project tree shows 'CLC_Data' expanded to 'reads'. The middle 'Toolbox' pane lists various tools, with 'Trim Sequences' highlighted. The right 'Trim Sequences' wizard window shows 'reads' selected in the 'Navigation Area' and 'Selected elements (*)' list.

- 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。

クオリティトリミング



Trim Sequences

1. Select sequencing data
2. Quality trimming

Set parameters

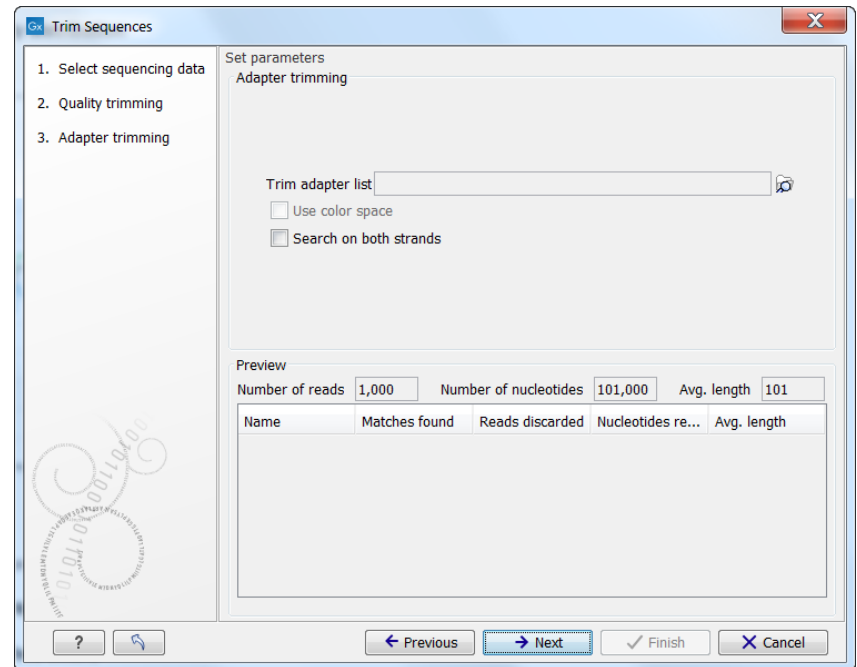
Quality trimming

Trim using quality scores
Limit: 0.05

Trim ambiguous nucleotides
Maximum number of ambiguities: 2

Previous Next Finish Cancel

- 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。



Trim Sequences

1. Select sequencing data
2. Quality trimming
3. Adapter trimming

Set parameters

Adapter trimming

Trim adapter list:

Use color space
 Search on both strands

Preview

Number of reads: 1,000 Number of nucleotides: 101,000 Avg. length: 101

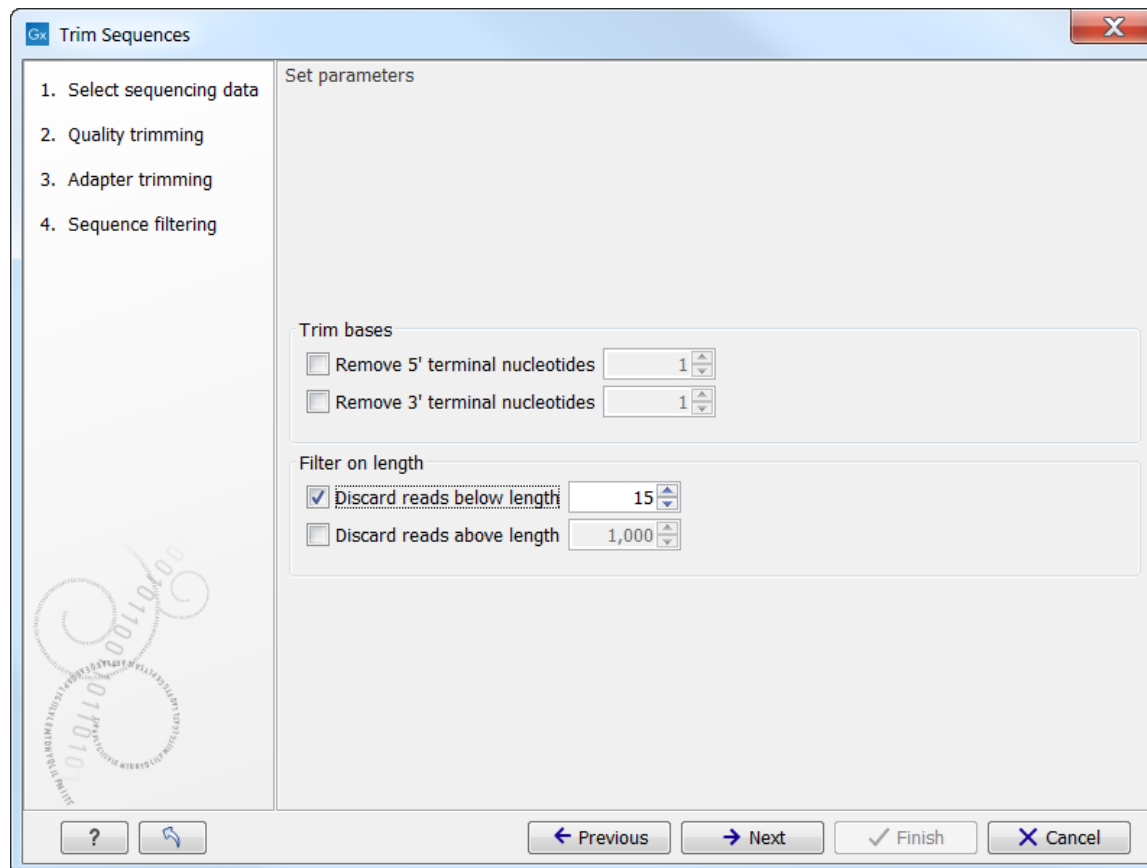
Name	Matches found	Reads discarded	Nucleotides re...	Avg. length

Previous Next Finish Cancel

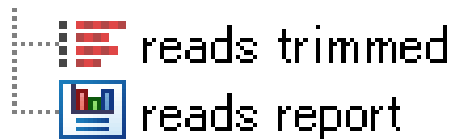
- 今回はアダプターは設定なし。

長さによるトリミング

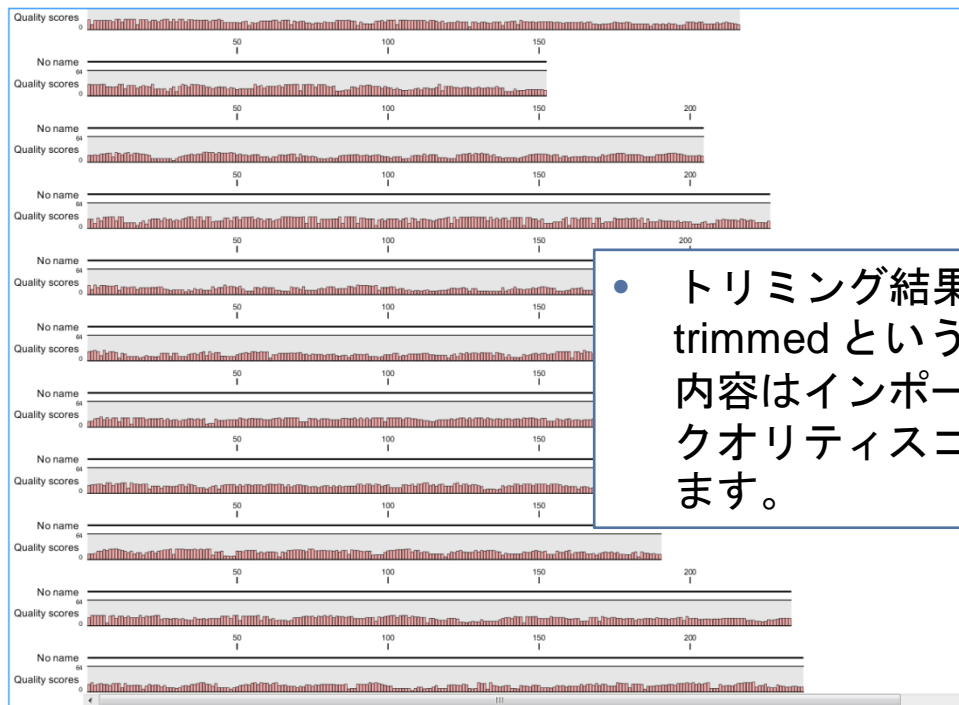
- 5末、3末の塩基数を指定してカットする
- Quality Scoreによるトリミングであまりに短いリードの除去など長さによるトリミング



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

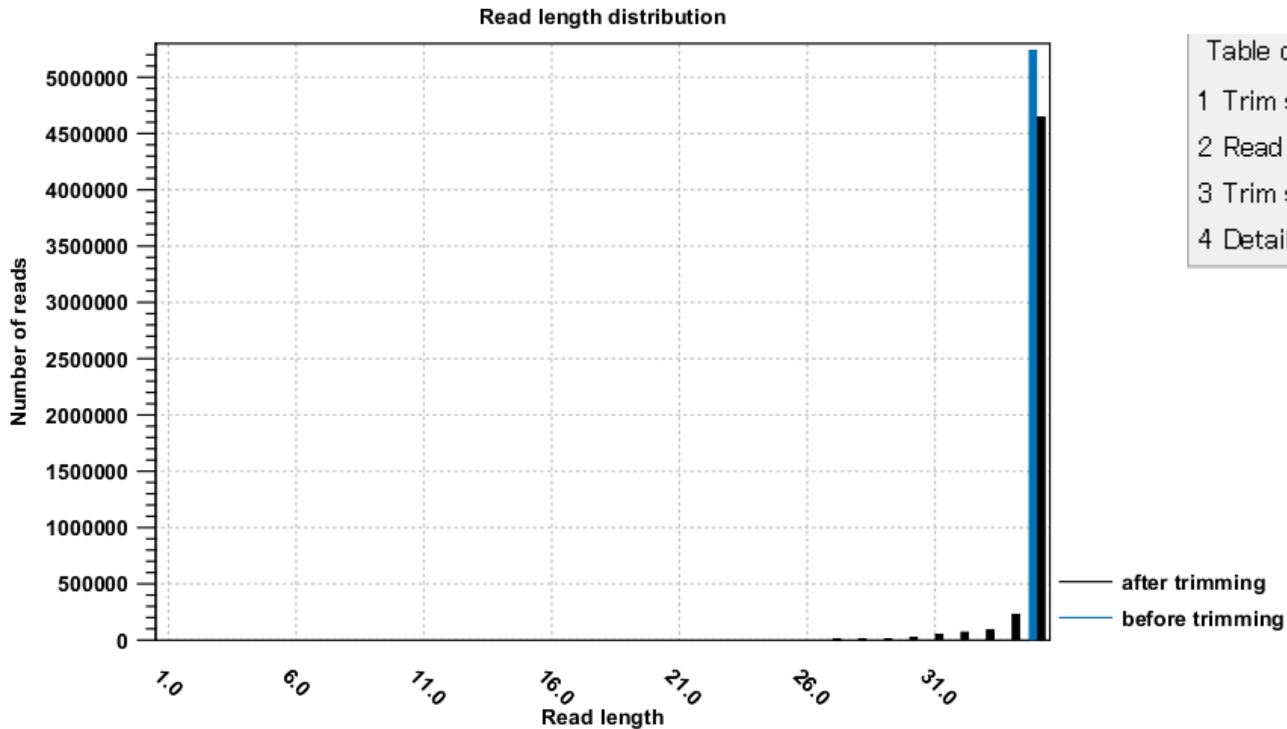


Table of Contents

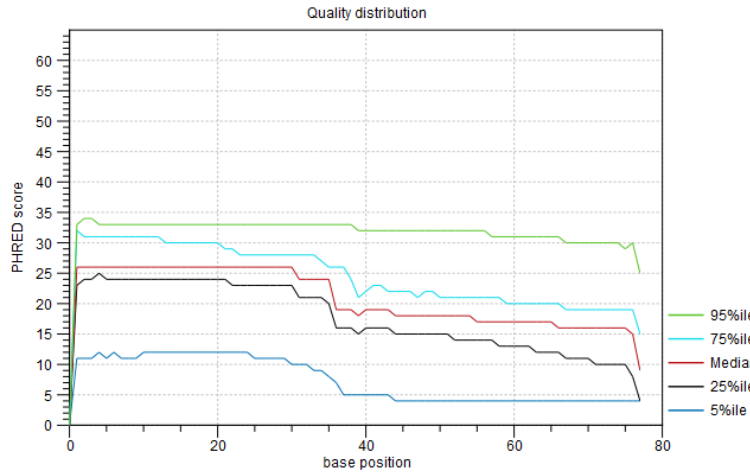
- 1 Trim summary
- 2 Read length before / after trimming
- 3 Trim settings
- 4 Detailed trim results

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

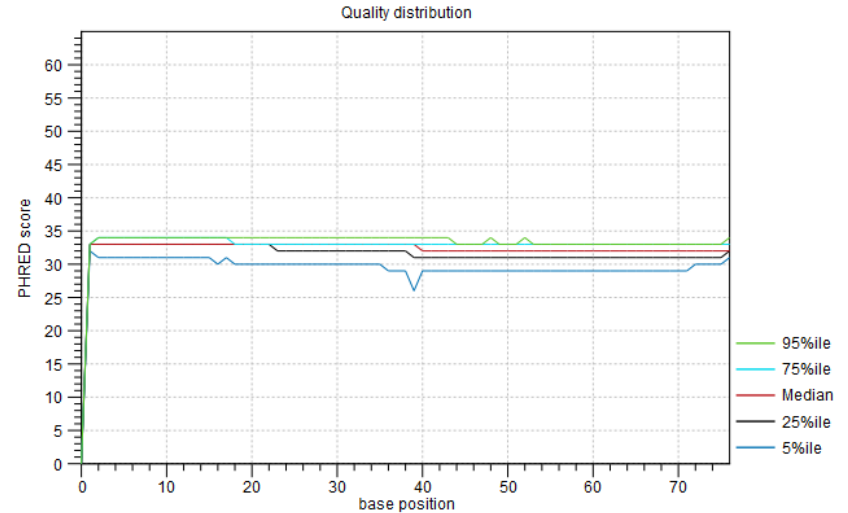
3.5 Quality distribution



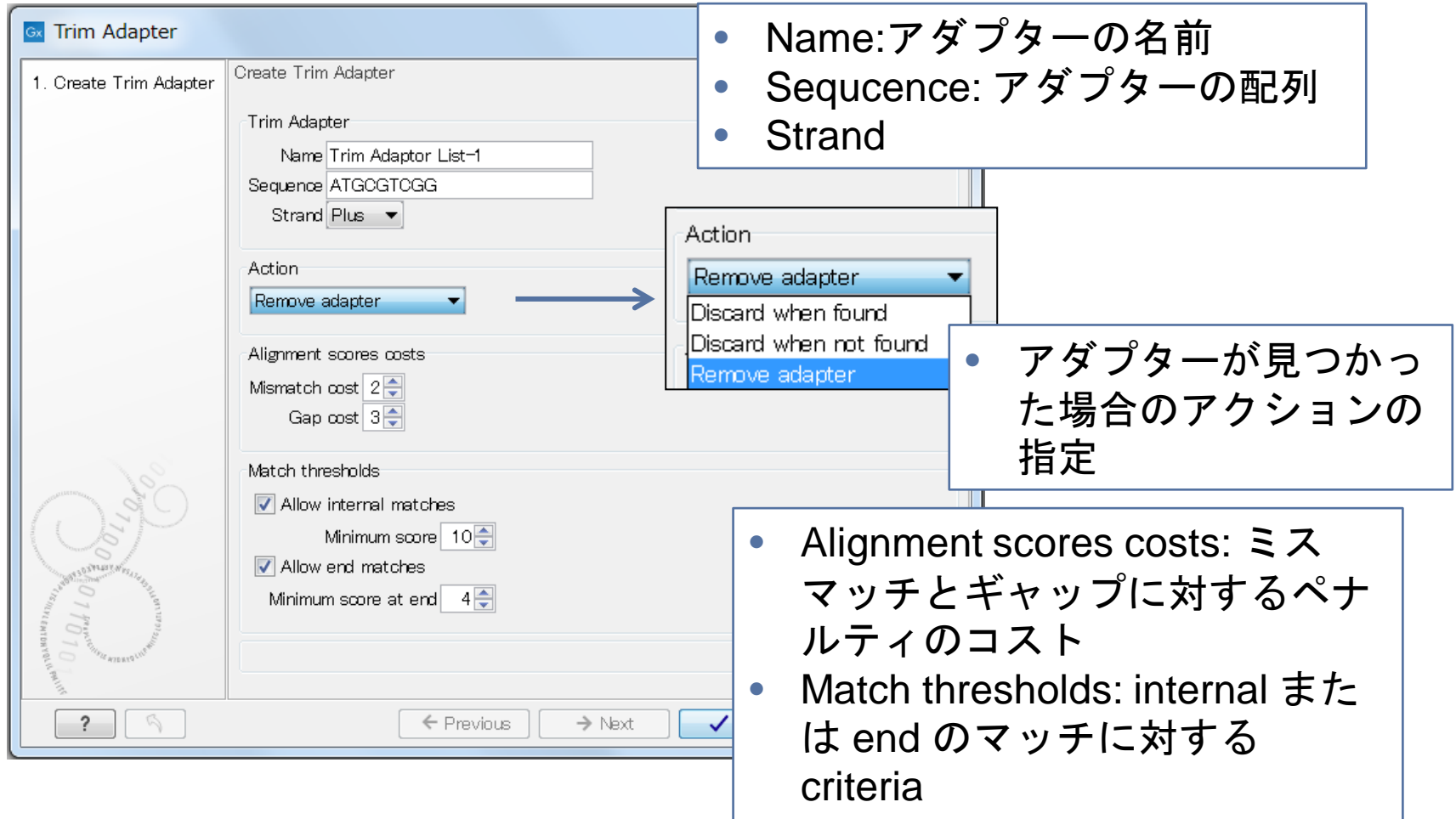
Base-quality distribution along the base positions.

After

3.5 Quality distribution



File > New > Trim Adapter List



- Name: アダプターの名前
- Sequence: 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

マッピング

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

```

CGTATCAATCGATTACGCTATGAATG
| | | | | | | | | | | | | | |
TTCAATCGATTACGCTATGA
  
```

アライメントスコア = 19

```

CGTATCAATCGATTACGCTATGAATG
| | | | | | | | | | | | | | |
TTCAATCAATTACGCTATGA
  
```

アライメントスコア = 16

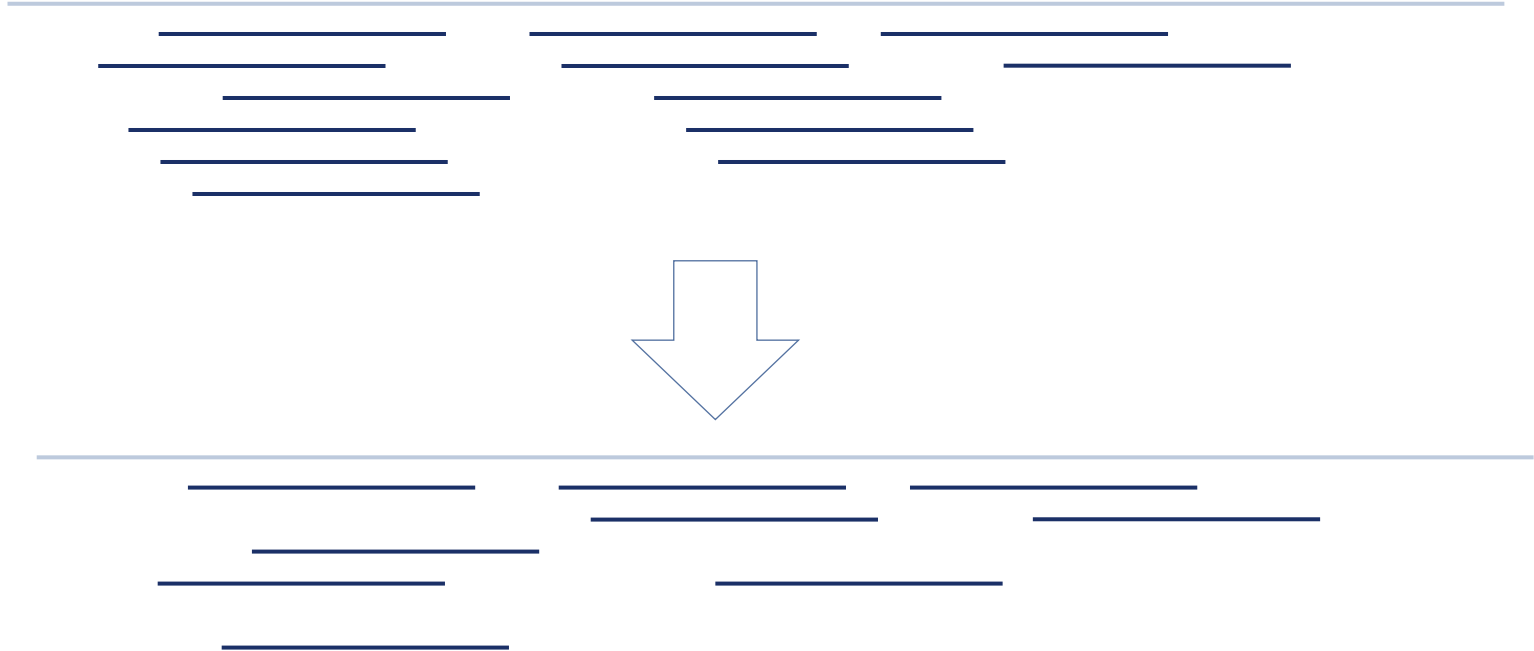
```

CGTATCAATCGATTACGCTATGAATG
| | | | | | | | | | | | | | |
TTCAATCAATTGCGCTATGC
  
```

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)

A

Genome: AATTCGCGCGGCATTCGCGCC

Read: AAATCG-----GCATTCGCGCC

50 match

$$50 + 6 + 4 \times (-3) + 11 = 55$$

B

Genome: AATTCGCGCGGCATTCGCGCC

Read: AAATCG-----GCATTCGCGCC

50 + 6 = 56

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

C

Genome: AATTCGCGCGGCATTCGCGCC

Read: AAATCG-----GCATTCGCGCC

$$50 + 6 + (-6) + 4 \times (-1) + 11 = 57$$

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。

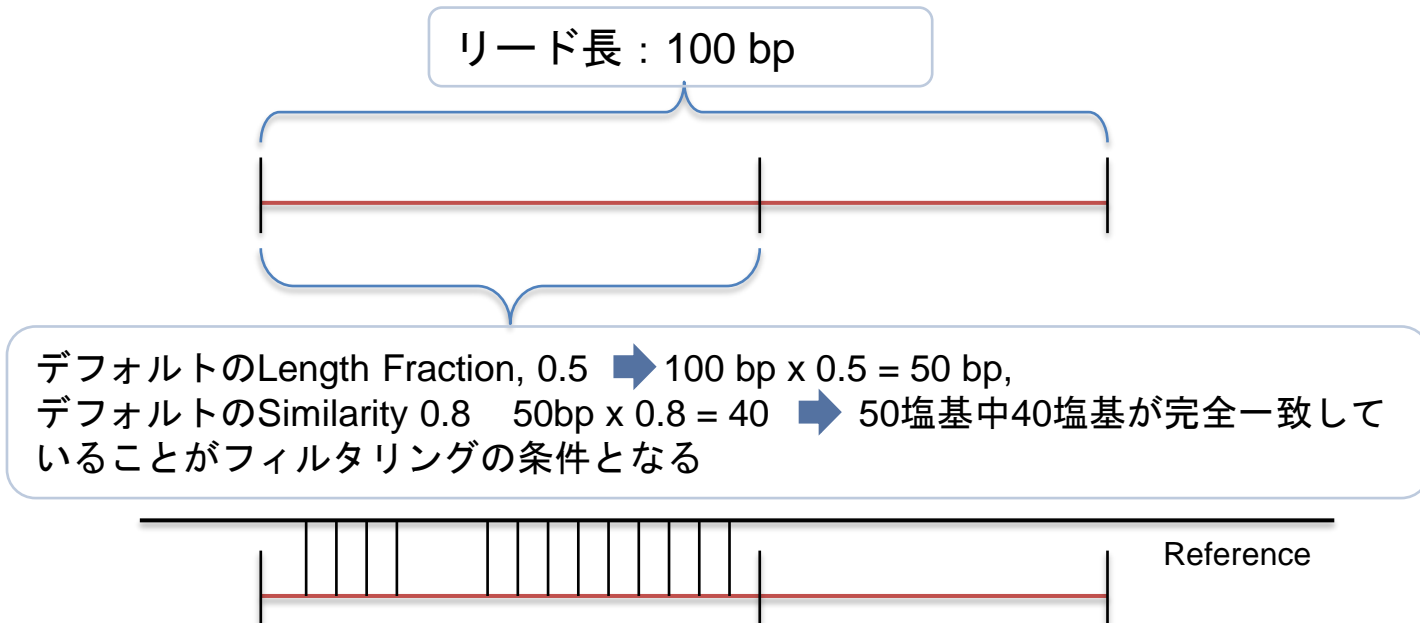
フィルタリング原理

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。

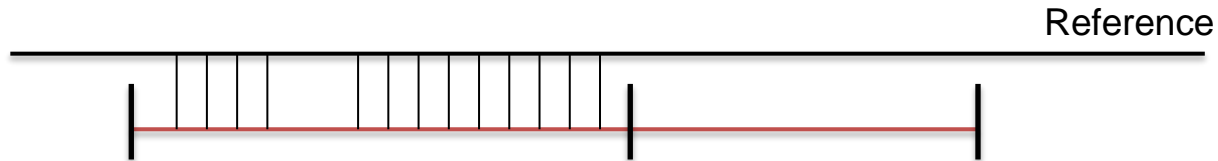
Length FractionとSimilarityは2つのパラメータの組み合わせで使用されます。

Length fraction: フィルターをかける際に、考慮する長さ

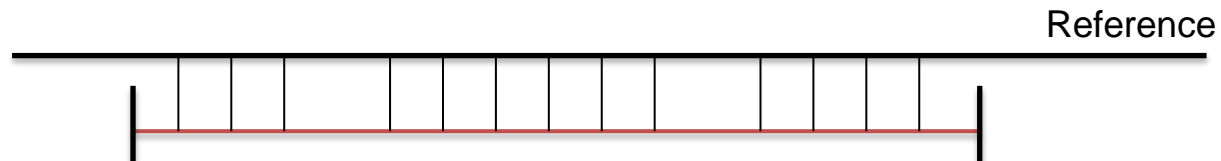
Similarity: Length Fractionで指定した長さのうち、どの程度類似しているものを残すか。



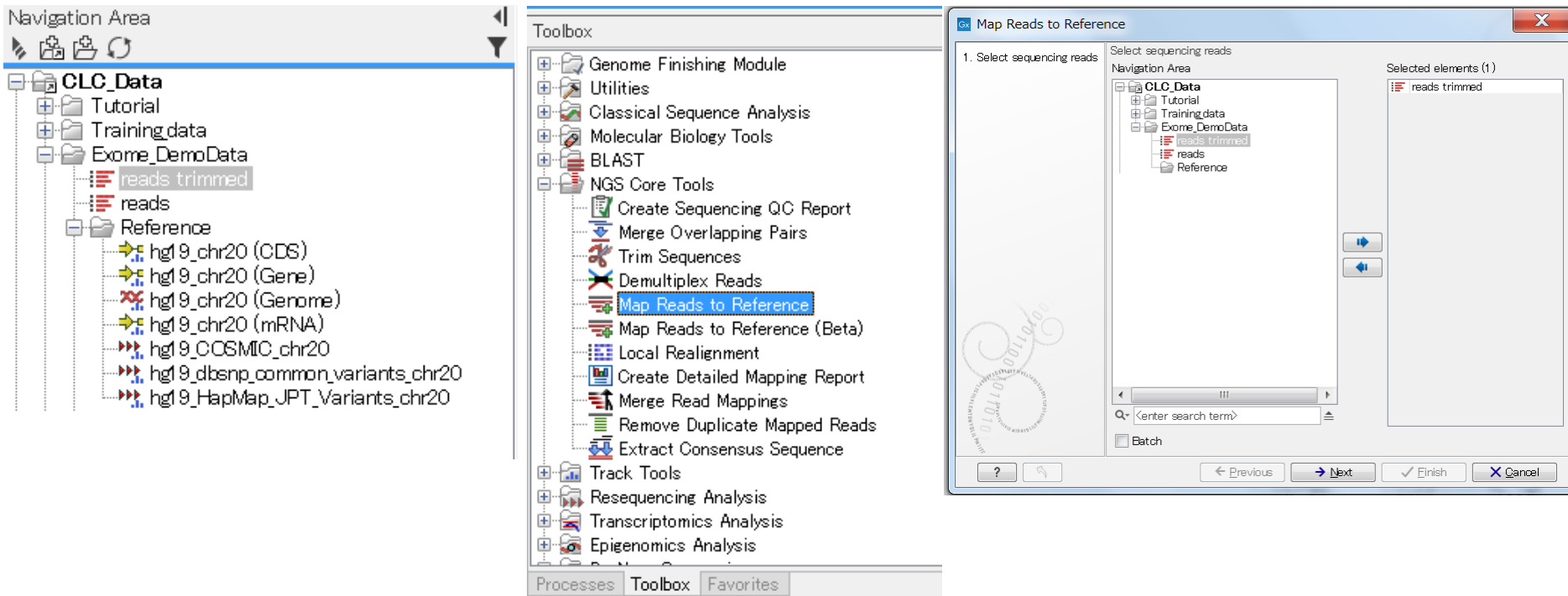
2つのパラメータを使う理由



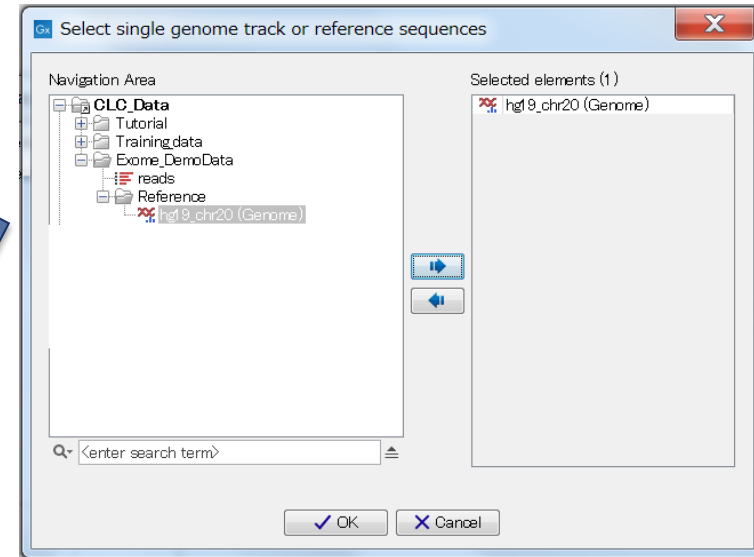
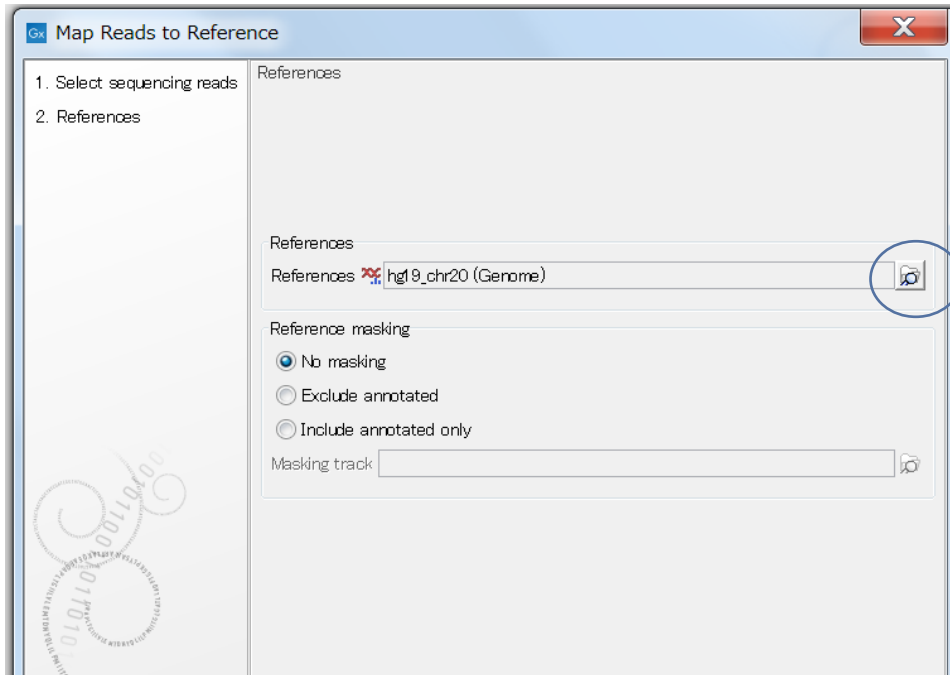
- リードの一部は似ているけれども、大きな挿入や、欠失によりリードの一部が参照配列と一致しない可能性がある場合
- トリミングが完全にできなかったクオリティの低い配列が末端部にある場合
(Length Fraction を小さくすることで、リードの一部に限定してアライメントの類似度を設定できる)



- 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



- 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。



- 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。

Gx Map Reads to Reference
⌵

1. Select sequencing reads
2. References
3. Mapping options

Mapping options

Read alignment

Mismatch cost

Mismatch の penalty

Linear gap cost

Affine gap cost

Insertion cost

Deletion cost

Insertion/deletion の penalty (Linear)

Insertion open cost

Insertion extend cost

Deletion open cost

Deletion extend cost

Length fraction

Similarity fraction

Filter の parameter

Global alignment

Color space alignment

Color error cost

Auto-detect paired distances

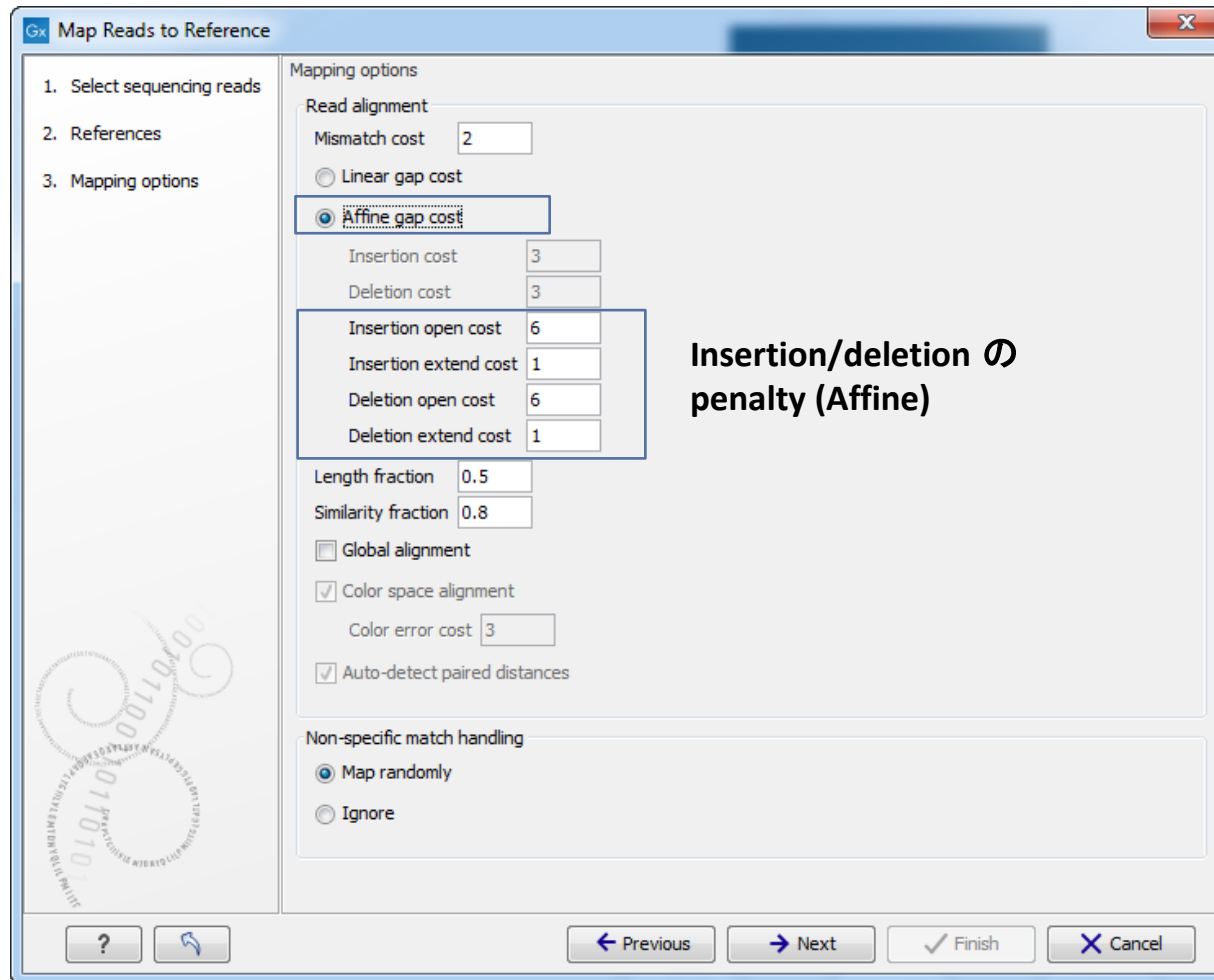
Non-specific match handling

Map randomly

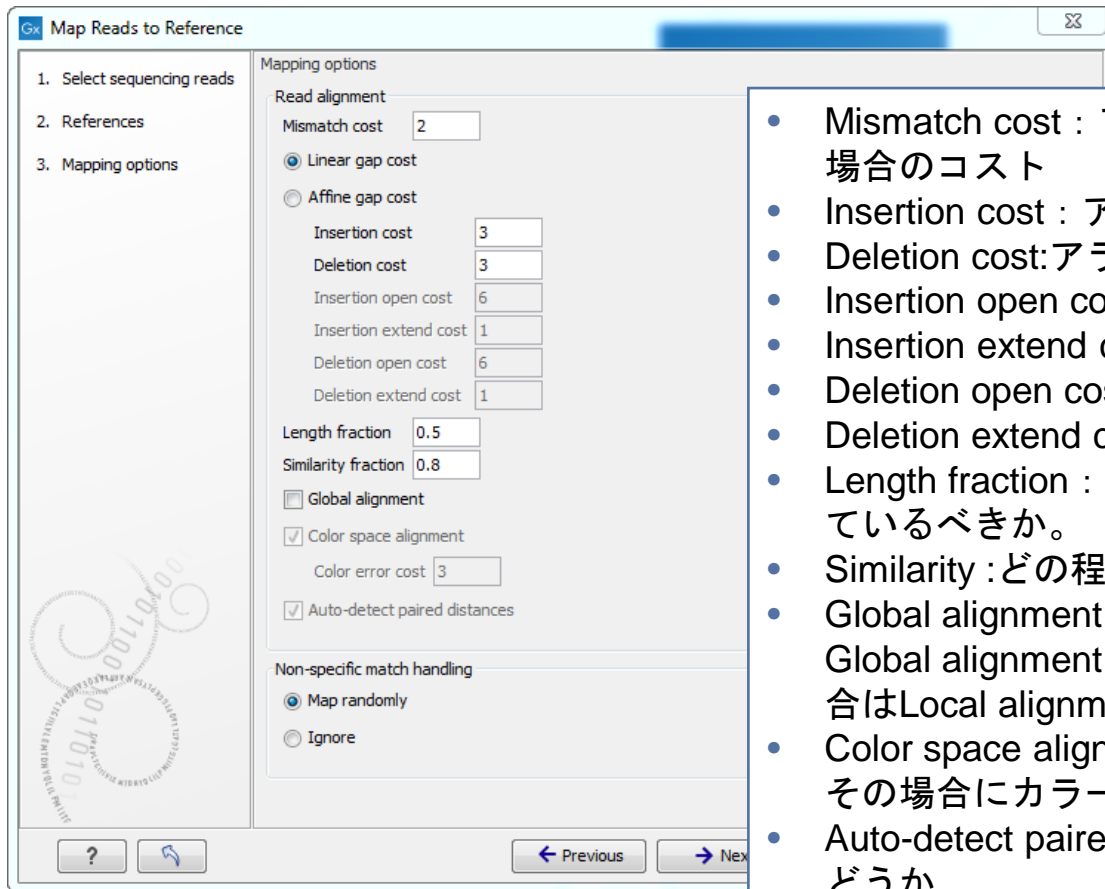
Ignore

?
↶

↷
✓ Finish
✕ Cancel

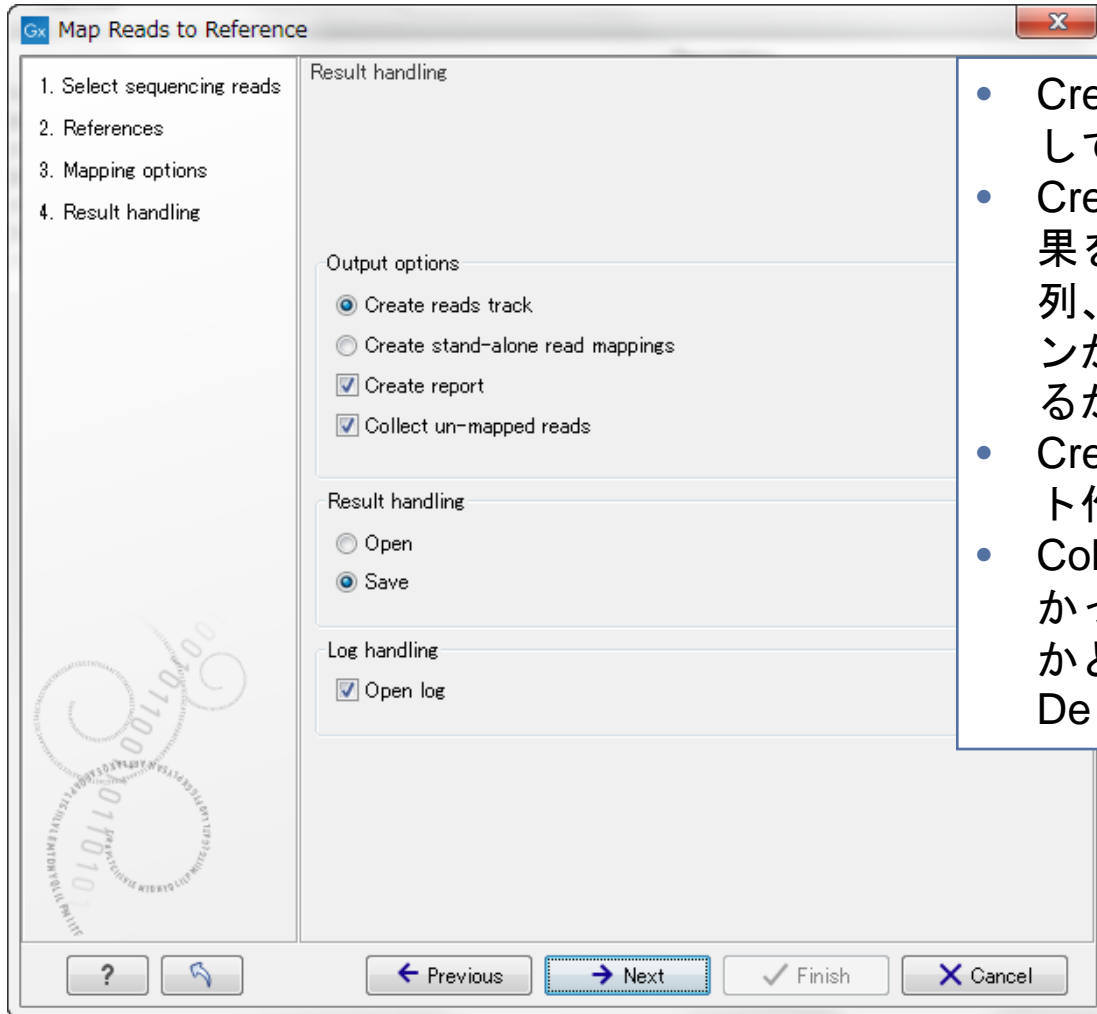


Insertion, Deletion 開始時点でカウントされるコスト (Insertion/Deletion open cost) と Insertion, Deletionが1塩基長くなる際に増えるコスト (Insertion/Deletion extended cost)



- Mismatch cost : アライメントにマッチしないものがあつた場合のコスト
- 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。

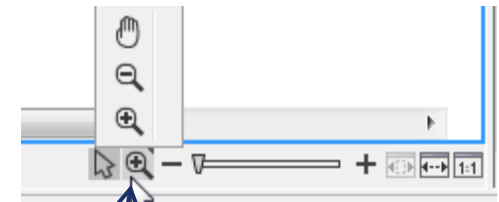
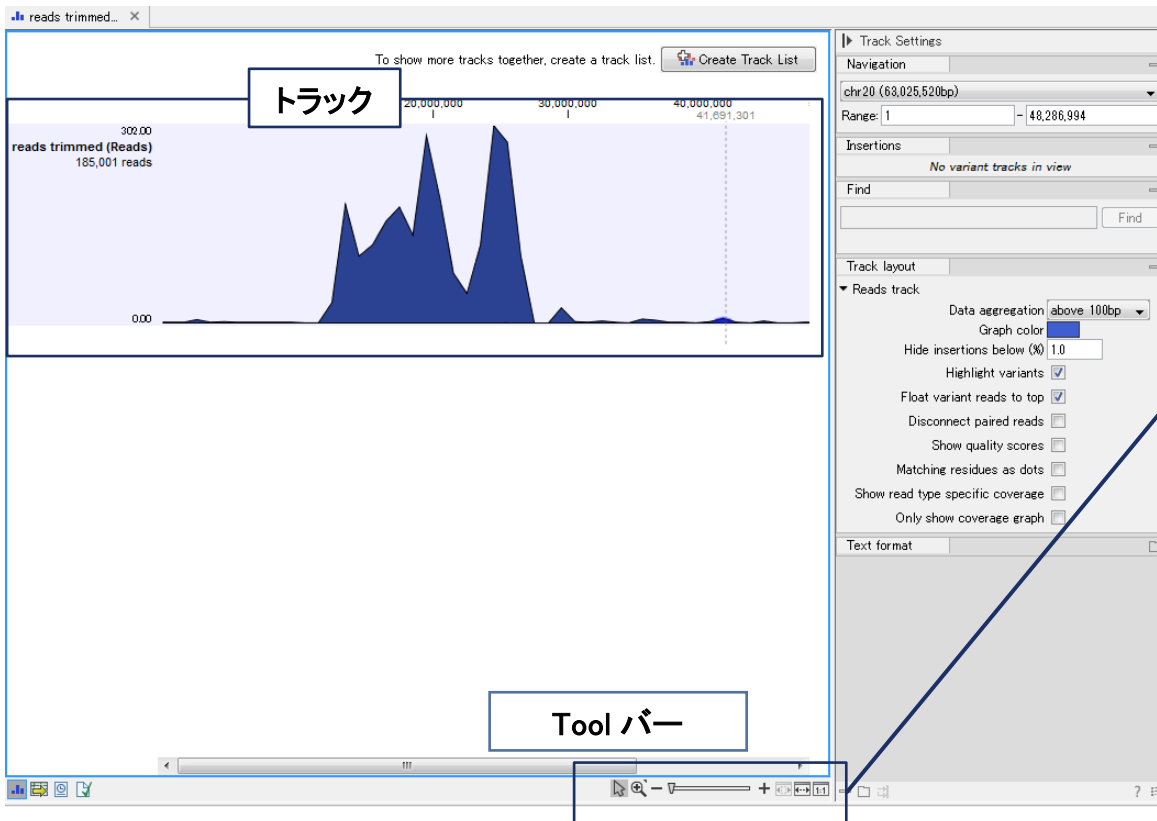
- Non-specific match handling : 同スコアでマップされる箇所がある場合の対処。



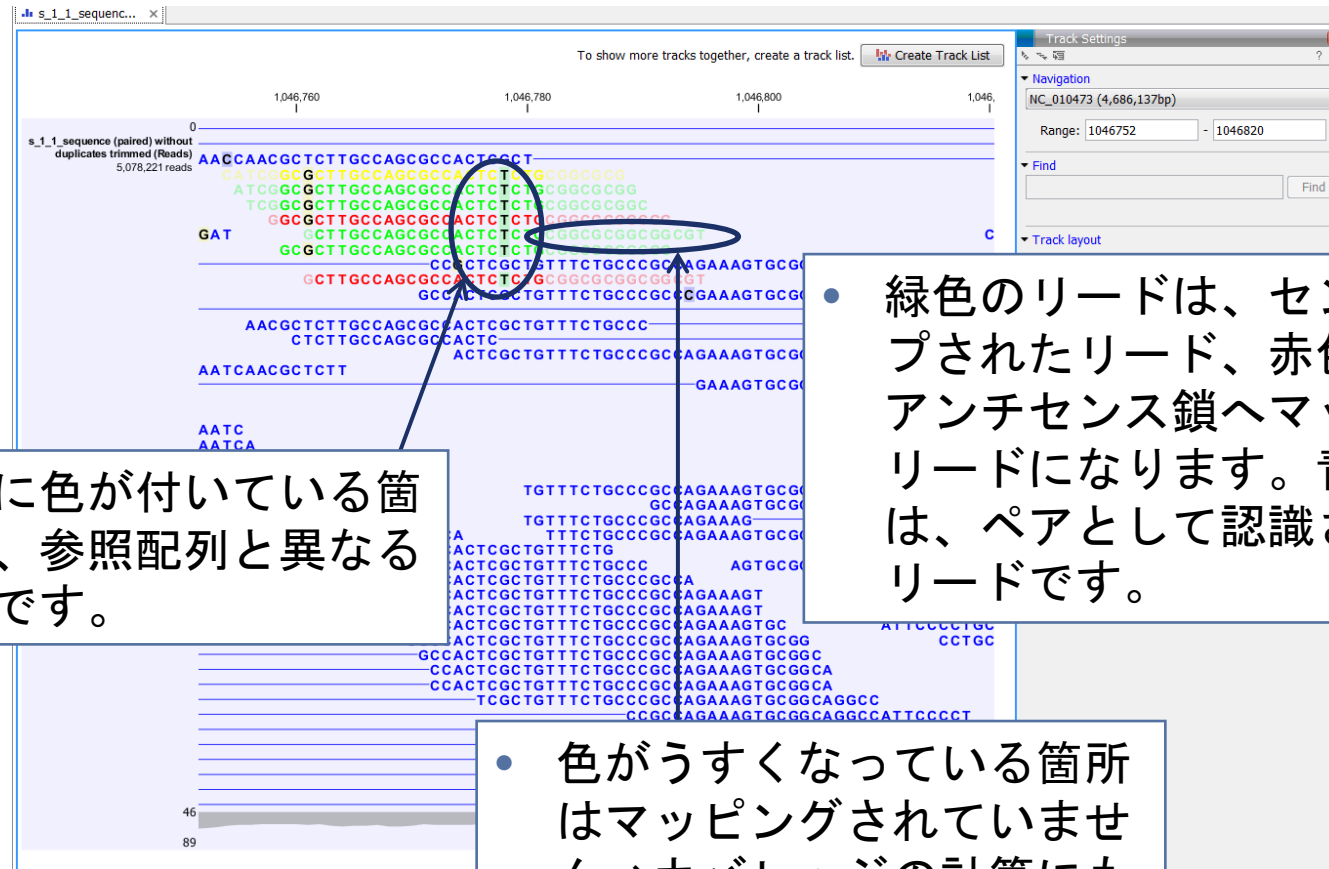
- Create reads track: 結果をトラックとして作成する場合。
- Create stand-alone read mappings : 結果をstand-aloneフォーマット（参照配列、リードマッピング、アノテーションが一つになったファイル）で作成するか。
- Create report: マッピング結果のレポート作成。
- Collect un-mapped reads: マップされなかったリードをリストとして作成するかどうか（リスト化することにより、De Novoなど、別の解析へ利用可能）

結果のトラック

- reads trimmed (Reads)
- reads trimmed un-mapped reads [no read group] (single)
- reads trimmed mapping summary report



- 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
- 1:1



- 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。

- 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。

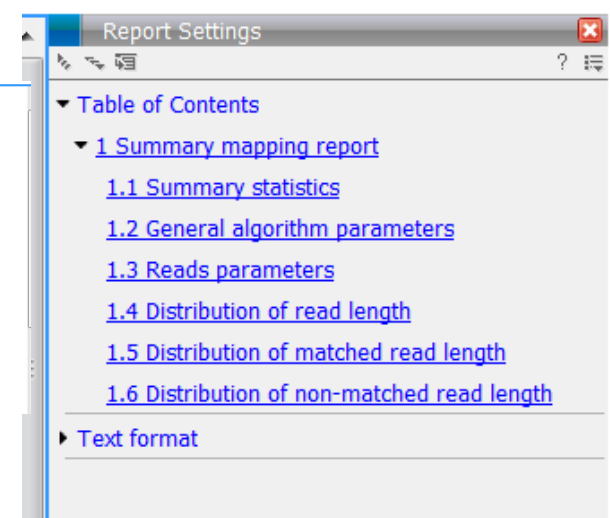
- 色がうすくなっている箇所はマッピングされていません⇒カバレッジの計算にも考慮されていません。

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

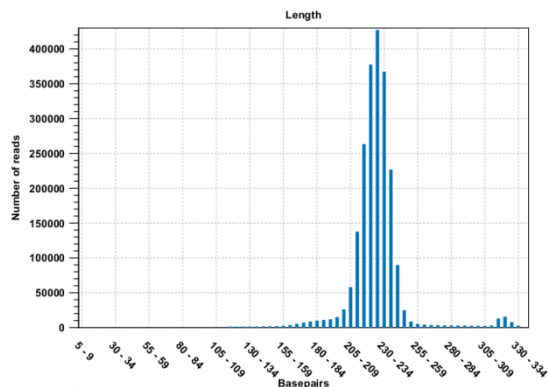
1 Summary mapping report

1.1 Summary statistics

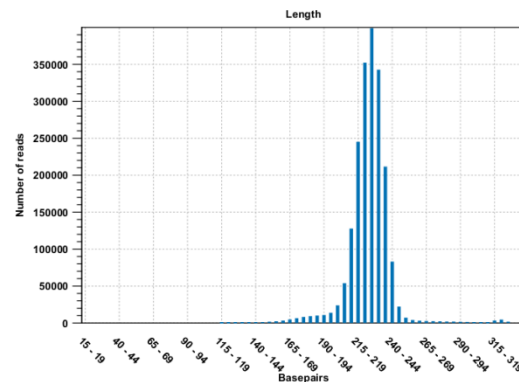
	Count	Percentage of reads	Average length	Number of bases	Percentage of bases
References	1	-	4,686,137.00	4,686,137	-
Mapped reads	5,078,221	97.95%	34.00	172,651,355	97.92%
Not mapped reads	106,167	2.05%	34.62	3,675,884	2.08%
Reads in pairs	4,987,262	96.20%	215.38	170,198,371	96.52%
Broken paired reads	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%



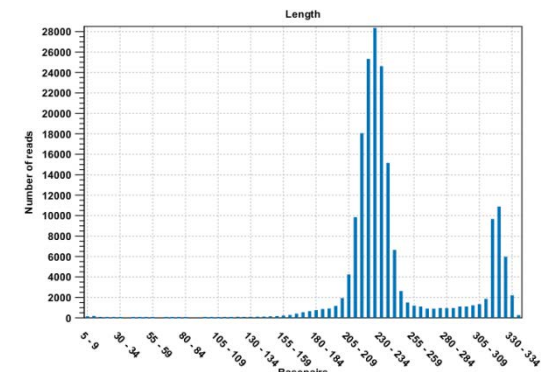
1.4 Distribution of read length



1.5 Distribution of matched read length

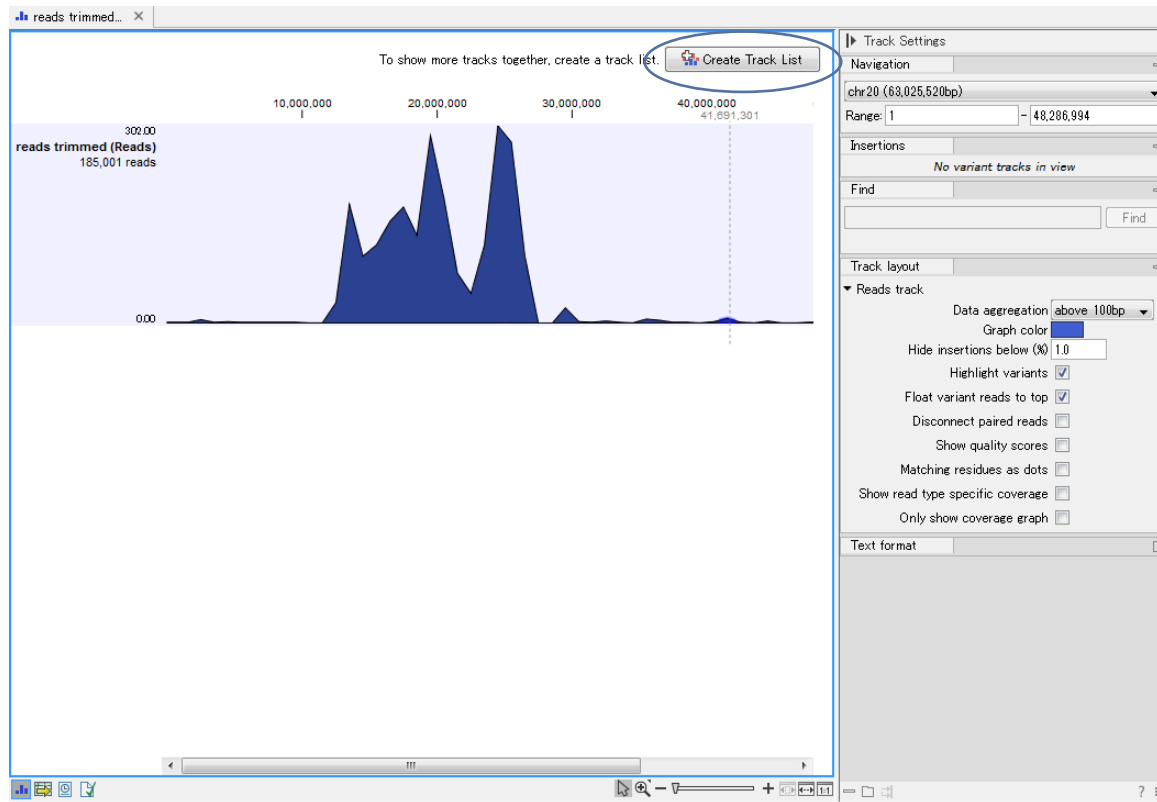


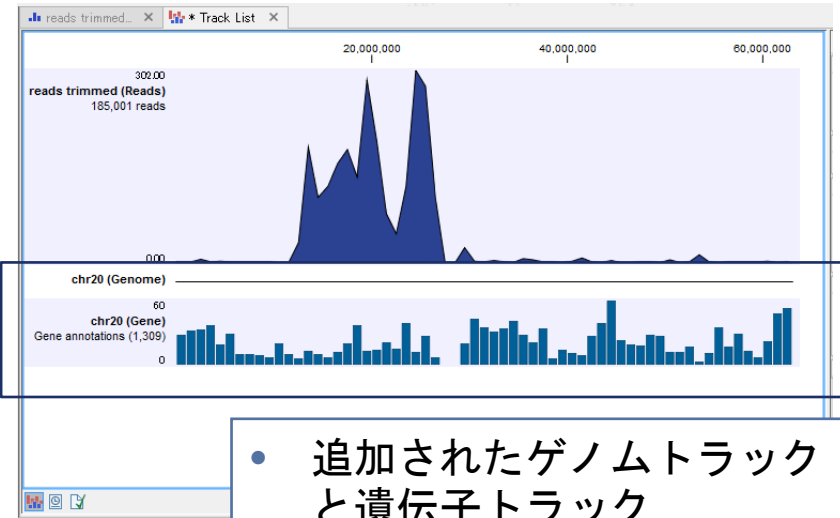
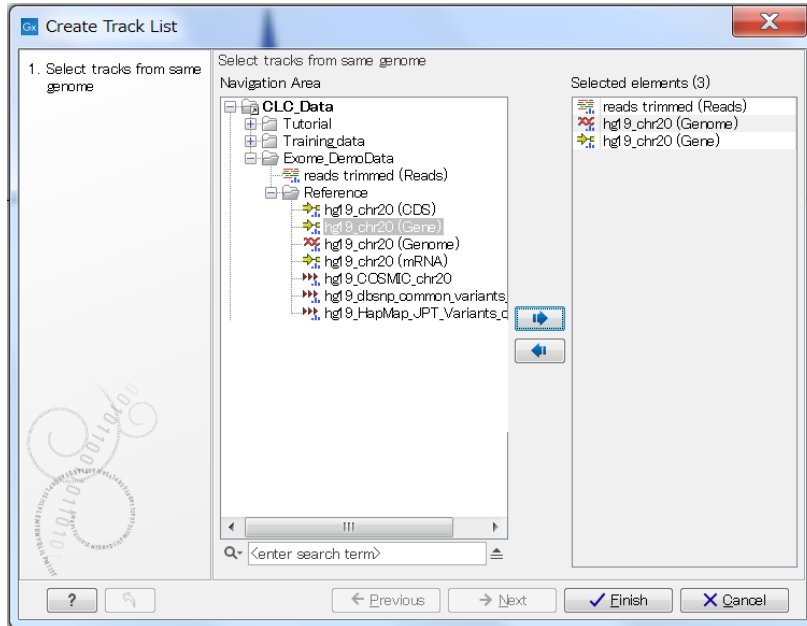
1.6 Distribution of non-matched read length



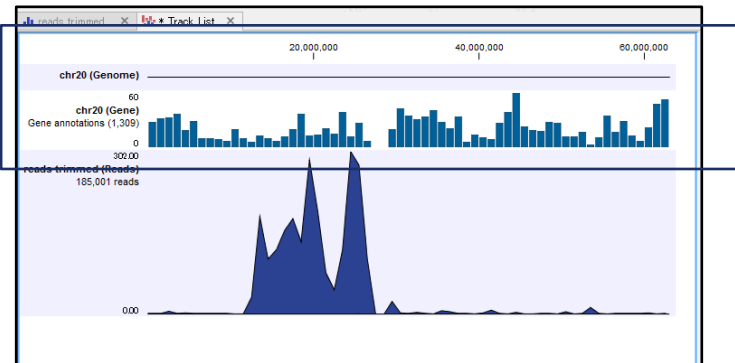
参照配列の追加

リードマッピングの結果に参照配列を追加しましょう。






● 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。

- 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。

Local Realignment

原理

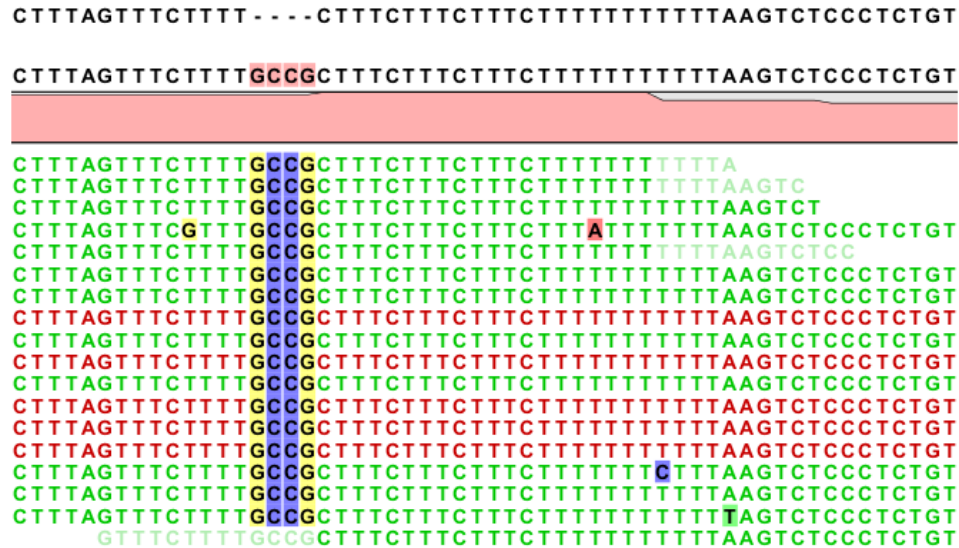
- Local Realignment では、このような状況を修正するため、マッピングを部分的にやり直します。この際、通常のマッピングの段階とは異なり、他のリードのマッピング状況を考慮するため、先ほどのマッピングは以下のように変化します。

```

CTTTAGTTTCCTTTT - - - CTTTCCTTCCTTTCTTTTTTTTTTTTTAAGTCTCCCTCTGT

CTTTAGTTTCCTTTT GCCGCTTTCCTTCCTTTCTTTTTTTTTTTTTAAGTCTCCCTCTGT

```

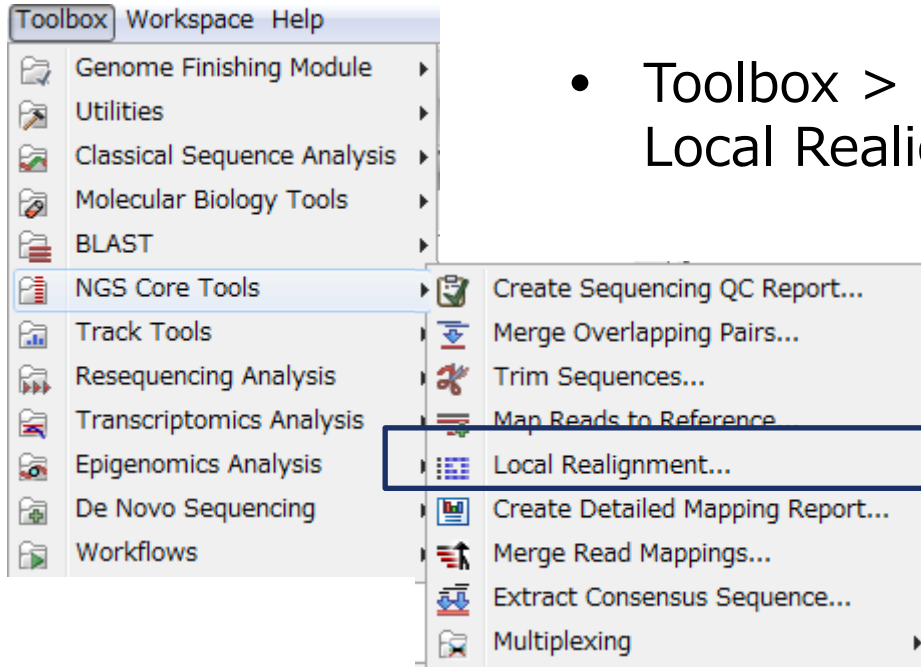


```

CTTTAGTTTCCTTTT GCCGCTTTCCTTCCTTTCTTTTTTTTTTTTTA
CTTTAGTTTCCTTTT GCCGCTTTCCTTCCTTTCTTTTTTTTTTTTTAAGTC
CTTTAGTTTCCTTTT GCCGCTTTCCTTCCTTTCTTTTTTTTTTTTTAAGTCT
CTTTAGTTTCCTTTT GTTTGCCGCTTTCCTTCCTTTCTTTTAAGTCTCCCTCTGT
CTTTAGTTTCCTTTT GCCGCTTTCCTTCCTTTCTTTTTTTTTTTTTAAGTCTCC
CTTTAGTTTCCTTTT GCCGCTTTCCTTCCTTTCTTTTTTTTTTTTTAAGTCTCCCTCTGT
CTTTAGTTTCCTTTT GCCGCTTTCCTTCCTTTCTTTTTTTTTTTTTAAGTCTCCCTCTGT
CTTTAGTTTCCTTTT GCCGCTTTCCTTCCTTTCTTTTTTTTTTTTTAAGTCTCCCTCTGT
CTTTAGTTTCCTTTT GCCGCTTTCCTTCCTTTCTTTTTTTTTTTTTAAGTCTCCCTCTGT
CTTTAGTTTCCTTTT GCCGCTTTCCTTCCTTTCTTTTTTTTTTTTTAAGTCTCCCTCTGT
CTTTAGTTTCCTTTT GCCGCTTTCCTTCCTTTCTTTTTTTTTTTTTAAGTCTCCCTCTGT
CTTTAGTTTCCTTTT GCCGCTTTCCTTCCTTTCTTTTTTTTTTTTTAAGTCTCCCTCTGT
CTTTAGTTTCCTTTT GCCGCTTTCCTTCCTTTCTTTTTTTTTTTTTAAGTCTCCCTCTGT
CTTTAGTTTCCTTTT GCCGCTTTCCTTCCTTTCTTTTTTTTTTTTTAAGTCTCCCTCTGT
CTTTAGTTTCCTTTT GCCGCTTTCCTTCCTTTCTTTTTTTTTTTTTAAGTCTCCCTCTGT
CTTTAGTTTCCTTTT GCCGCTTTCCTTCCTTTCTTTTTTTTTTTTTAAGTCTCCCTCTGT
CTTTAGTTTCCTTTT GTTTGCCGCTTTCCTTCCTTTCTTTTTTTTTTTTTAAGTCTCCCTCTGT

```

- 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。



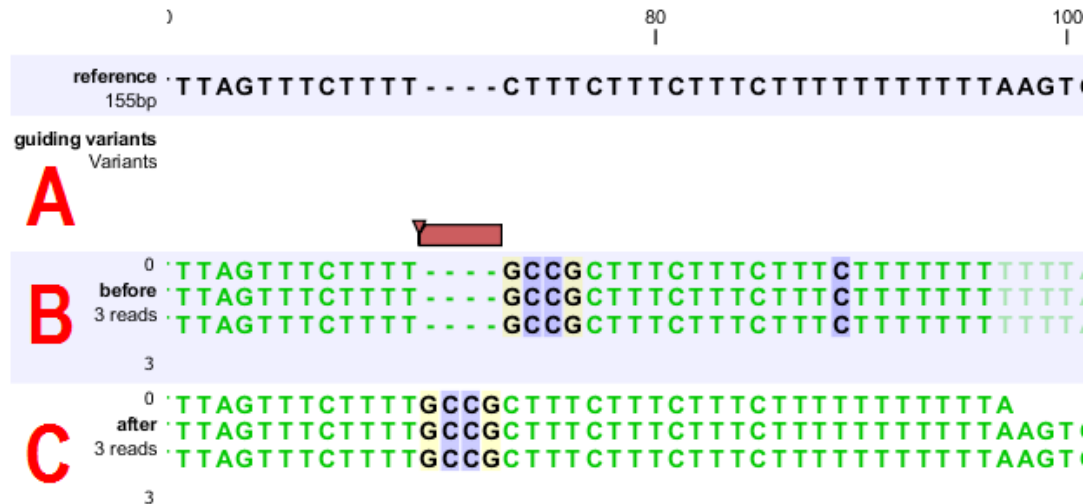
- Toolbox > NGS Core Tools > Local Realignment

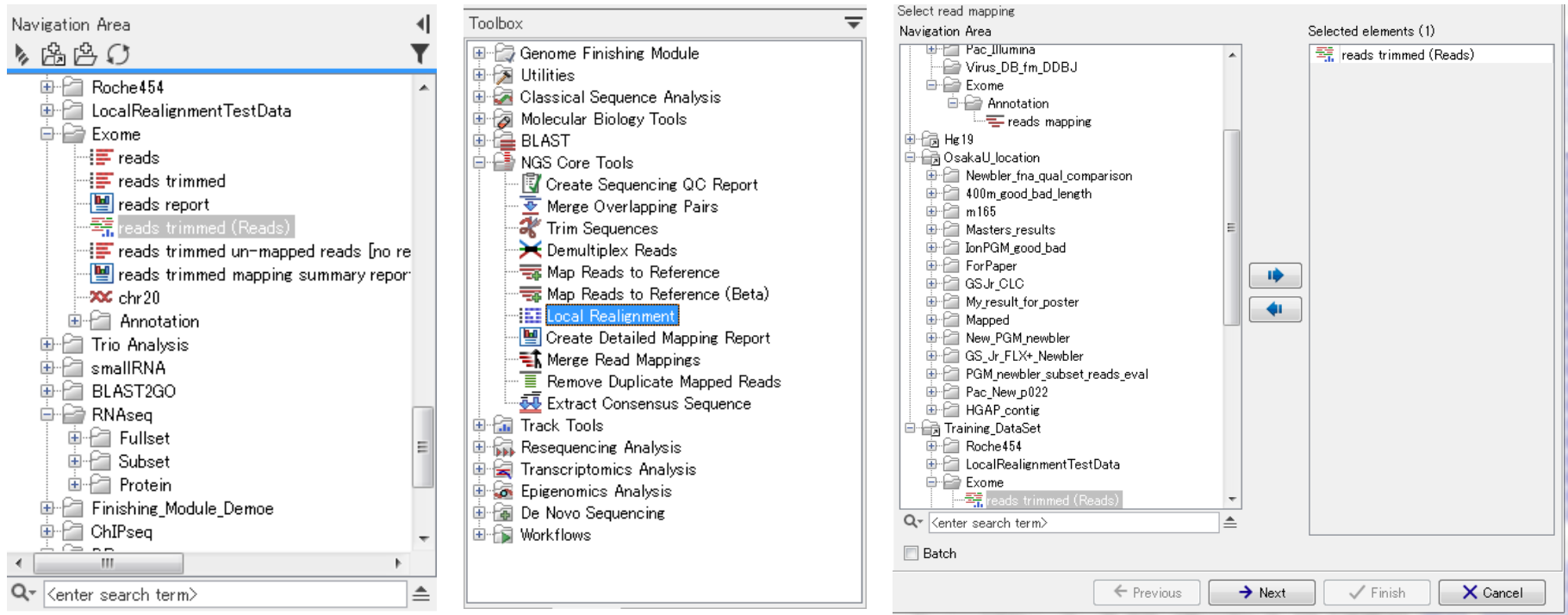
2種類のLocal Realignmentsがあります。さらにGuided にNo forceとForceの2種類があります。

- Non guided
- Guided
 - No force
 - Force

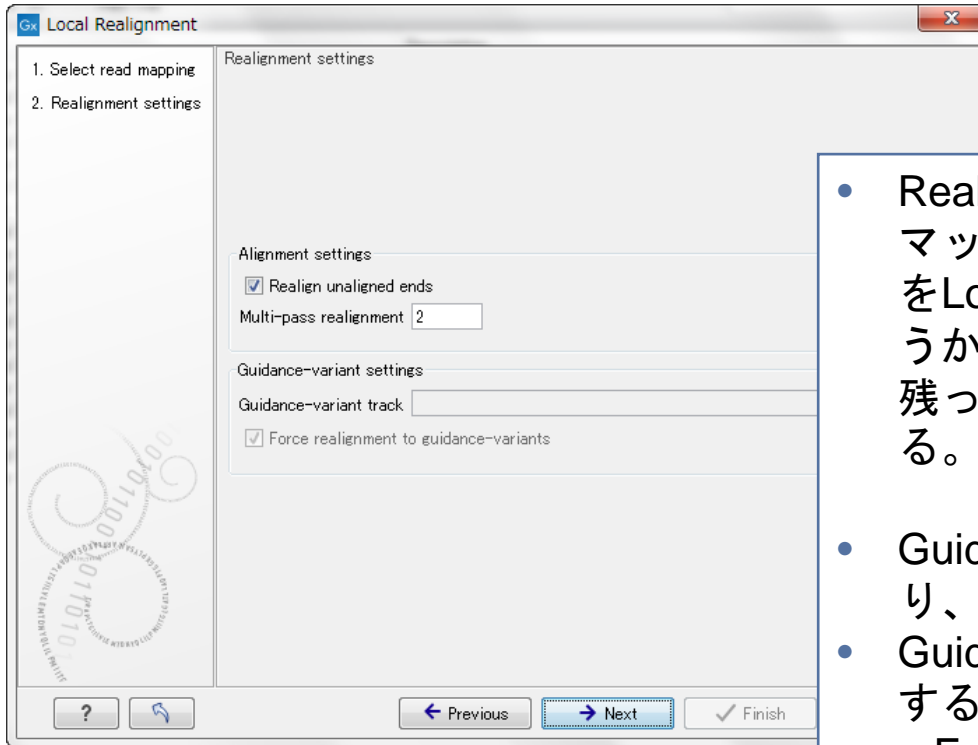
Guided Local Realignment

- 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**

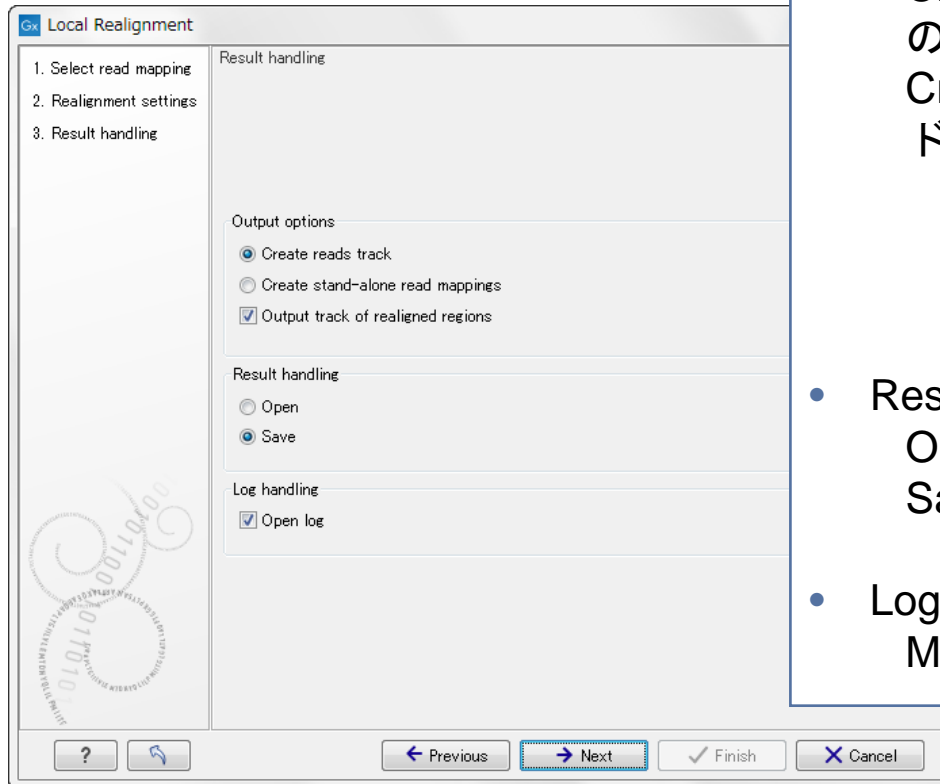




- Navigation Areaから使用するマッピングデータを選択。
- Toolboxから NGS Core Tools > Local Realignment を選択、ダブルクリック。
- ウィザードが起動し、選択したデータが選ばれていることを確認。



- 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。



- 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。

- 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

A screenshot of a genomic browser showing a track for 'chr2 Big selection (Reads) - locally realigned'. The track displays a series of horizontal bars representing reads, with a vertical dashed line indicating a specific genomic position.

- トラックフォーマットで作成した場合

A screenshot of a genomic browser showing a track for 'chr2 Big selection (Reads) - locally realigned'. The track displays a series of horizontal bars representing reads, with a vertical dashed line indicating a specific genomic position.

- 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

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

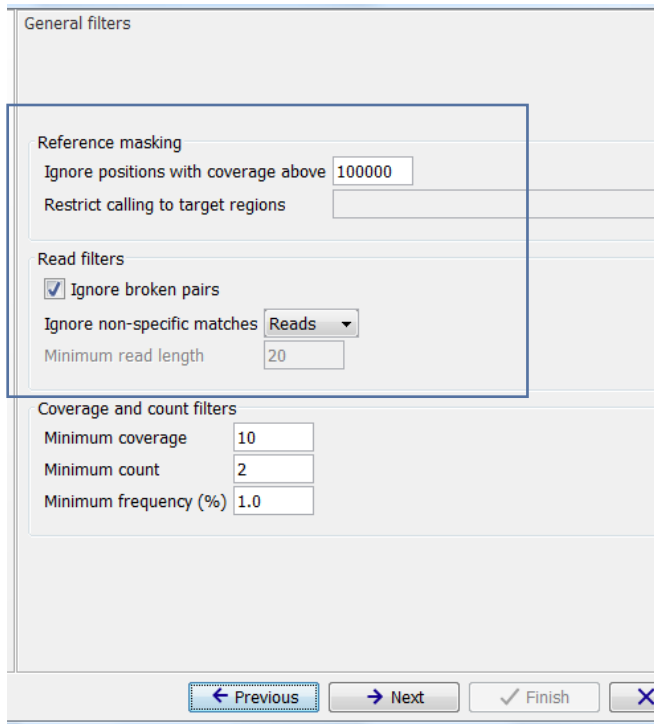
(version 7.5以前のProbabilistic Variant Detection)

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。

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エーションについてリストに含めない

Restrict calling to target regions : バリエーション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エーション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、その領域のバリエーションを検出しません。
- 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、その長さからnon-specificになる可能性があるためです。

共通フィルター

General filters

Reference masking

Ignore positions with coverage above

Restrict calling to target regions

Read filters

Ignore broken pairs

Ignore non-specific matches

Minimum read length

Coverage and count filters

Minimum coverage

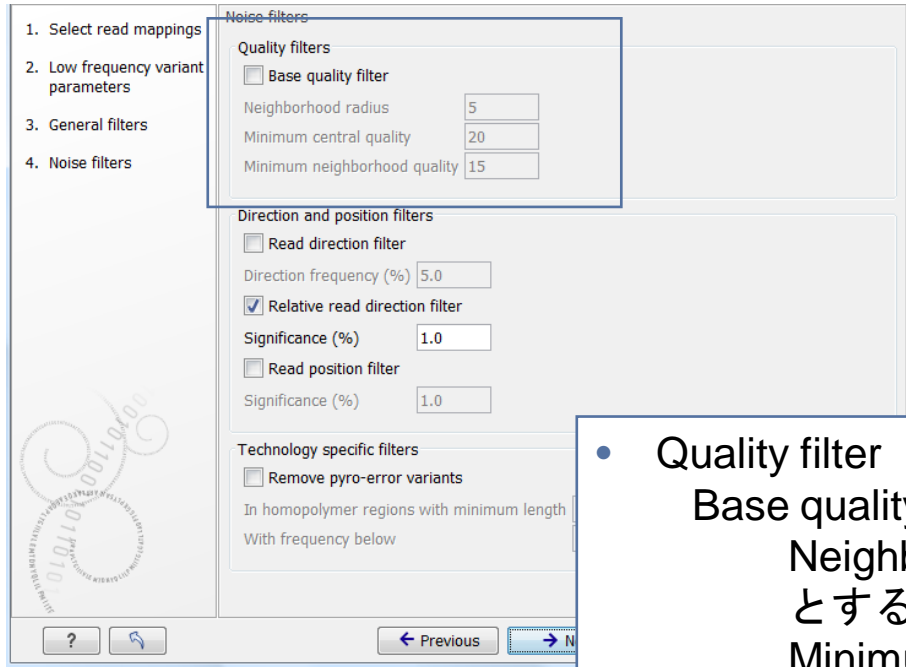
Minimum count

Minimum frequency (%)

Coverage and count filters

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

共通フィルター

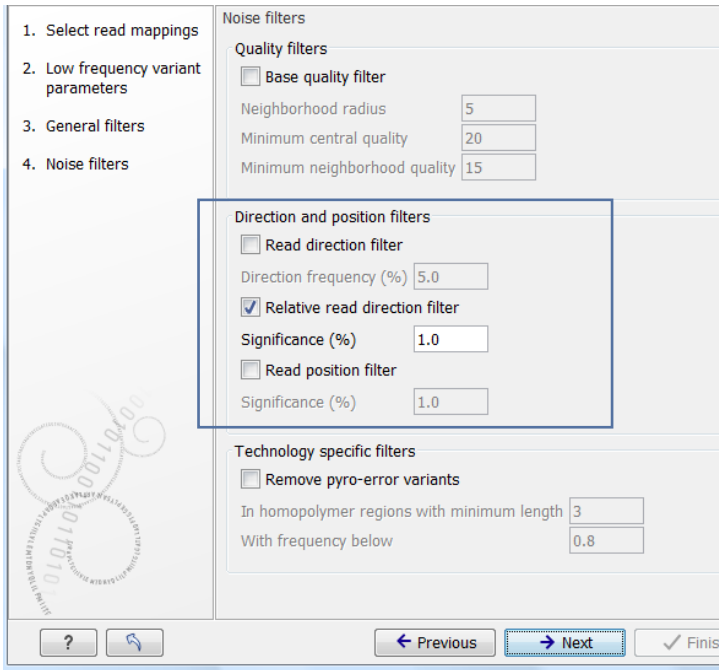


```
>TTTTTGCACCTCATTTCATATAAAAAATATATTTCCCCACG
>TTTTTGCACCTCATTTCATATAAAAAATATATTTCCCCACG
>TTTTTGCACCTCATTTCATATAAAATAAATATATTTCCCCACG
>TTTTTGCACCTCATTTCATATAAAAAATATATTTCCCCACG
>TTTTTGCACCTCATTTCATATAAAAAATATATTTCCCCACG
>TTTTTGCACCTCATTTCATATAAAAAATATATTTCCCCACG
>TTTTTGCACCTCATTTCATATAAAAAATATATTTCCCCACG
>TTTTTGCACCTCATTTCATATAAAAAATATATTTCCCCACG
>TTTTTGCACCTCATTTCATATAAAAAATATATTTCCCCACG
|ACTCATTTCATATAAAAAATATATTTCCCCACG
|CTCATTTCATATAAAAAATATATTTCCCCACG
|ATAAAAAATATATTTCCCCACG
|CCACG
```

- 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)

共通フィルター



1. Select read mappings

2. Low frequency variant parameters

3. General filters

4. Noise filters

Noise filters

Quality filters

Base quality filter

Neighborhood radius

Minimum central quality

Minimum neighborhood quality

Direction and position filters

Read direction filter

Direction frequency (%)

Relative read direction filter

Significance (%)

Read position filter

Significance (%)

Technology specific filters

Remove pyro-error variants

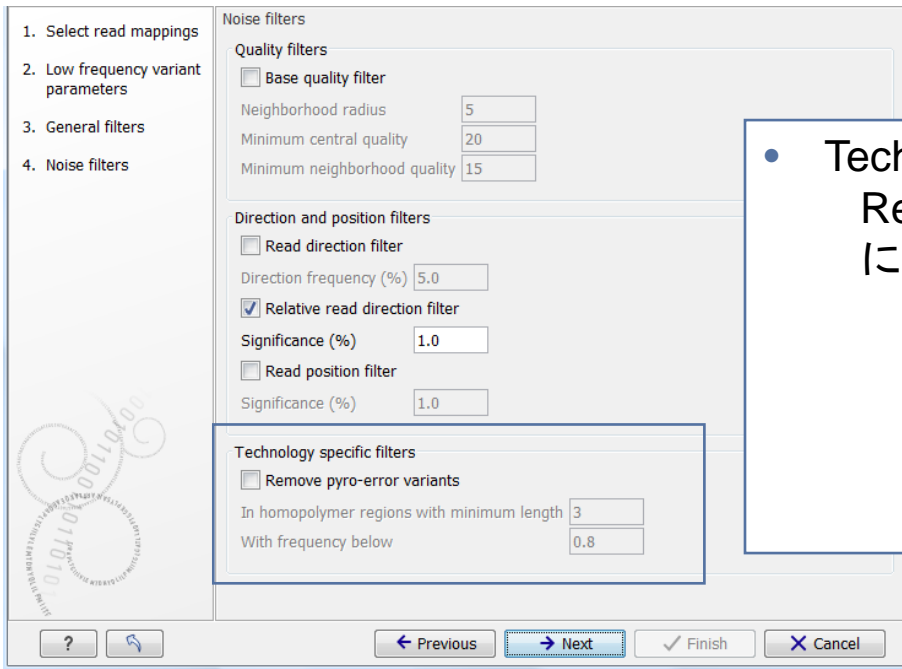
In homopolymer regions with minimum length

With frequency below

? ↶ ↷ ✓ Finish

- 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 : システム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、バリエーション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

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。

フィルター例

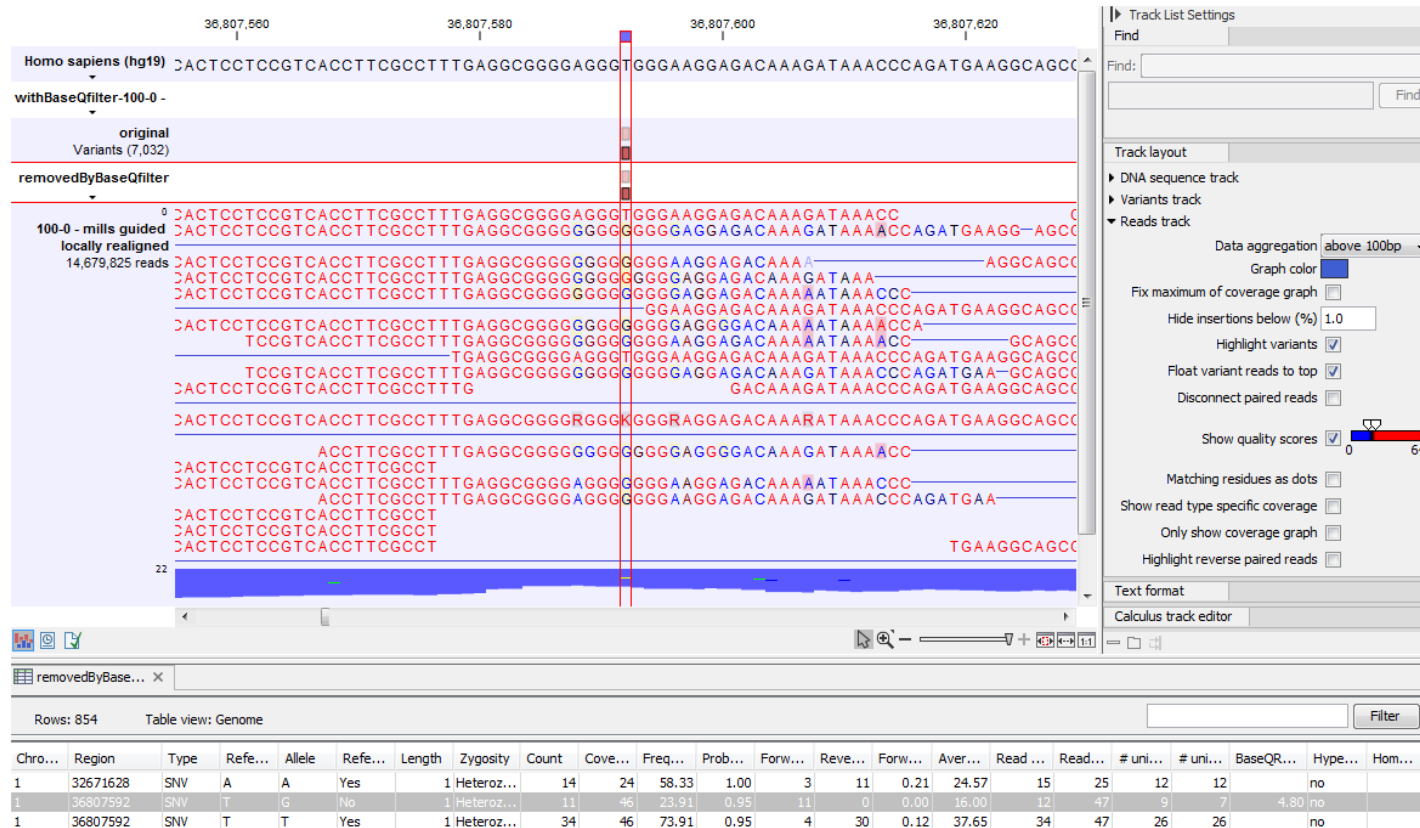
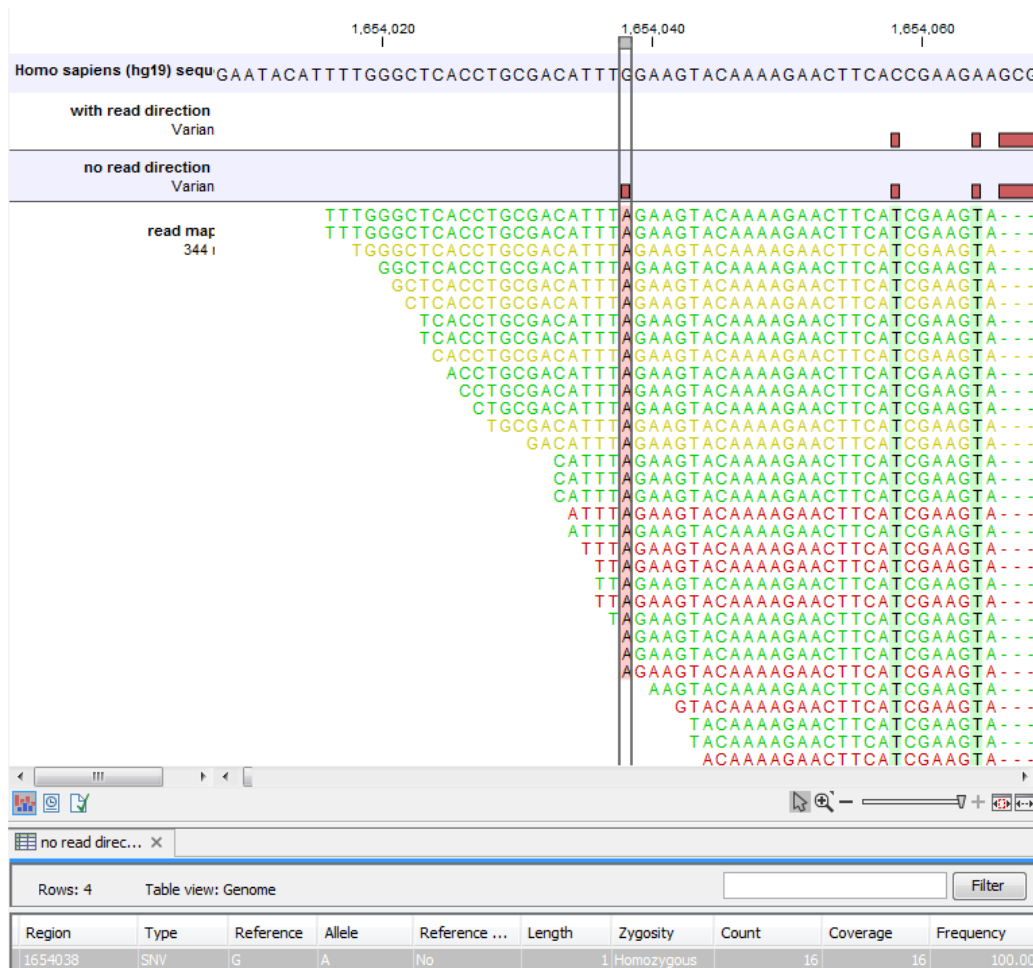


Figure 27.22: The same data as in figure 27.21, now with the 'Show quality scores' option in the reads track switched on.

- 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エーションのリストからはずされます。

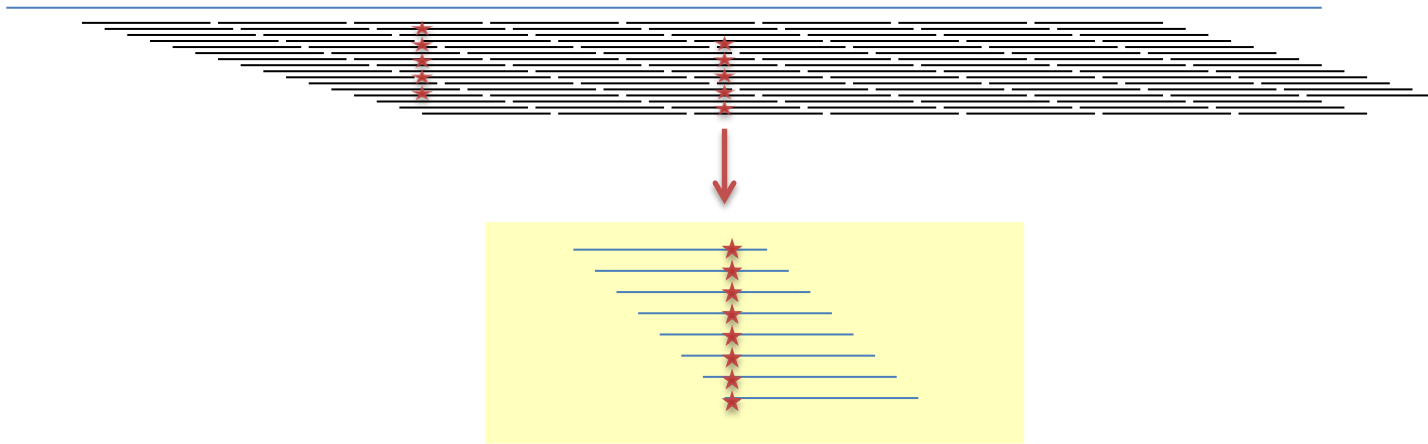
フィルター例



- 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エーション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エーションが取り除かれる。

原理

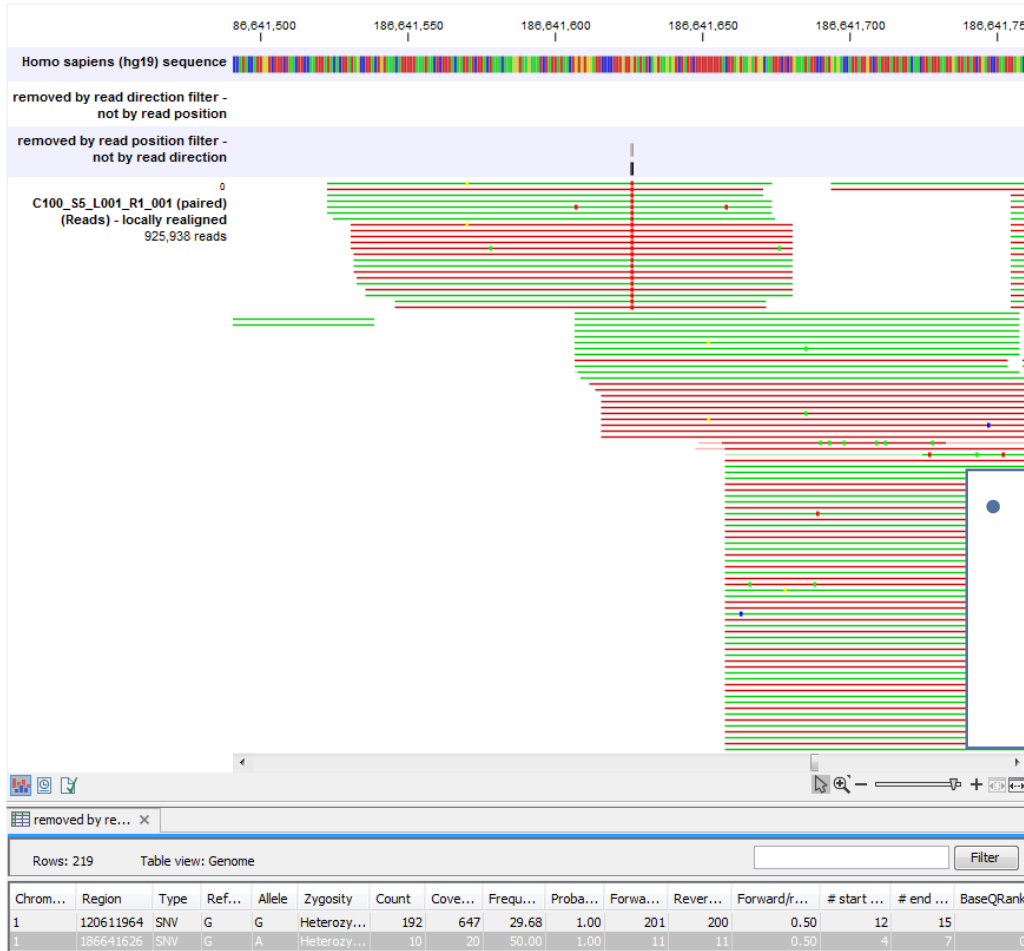
- もしリードが理想的な均一なカバレッジであれば、検出されるバリエーションをサポートする塩基のリード中の位置は、さまざまになるはずですが。



- 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



フィルター一例



- Read position filter 適用例：バリエーション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、このバリエーションは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

1. Select read mappings
2. Low frequency variant parameters
3. General filters
4. Noise filters

Noise filters

Quality filters

Base quality filter

Neighborhood radius

Minimum central quality

Minimum neighborhood quality

Direction and position filters

Read direction filter

Direction frequency (%)

Relative read direction filter

Significance (%)

Read position filter

Significance (%)

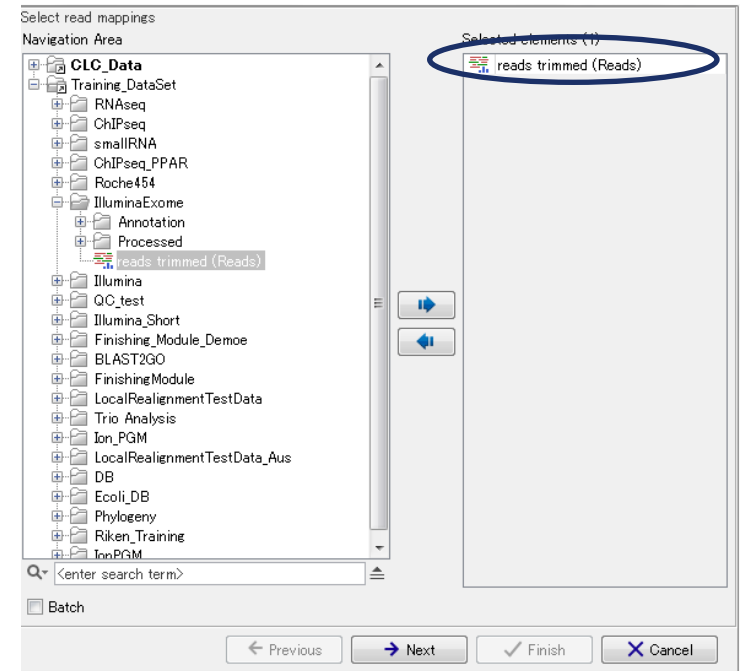
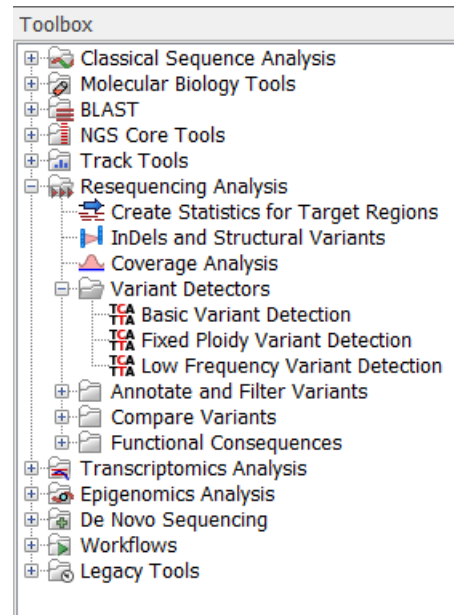
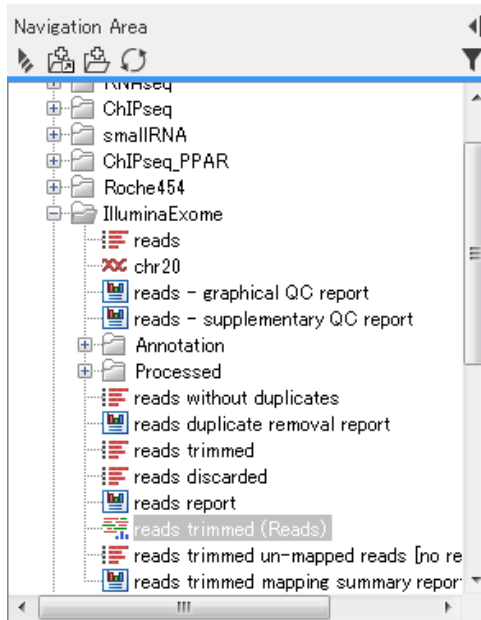
Technology specific filters

Remove pyro-error variants

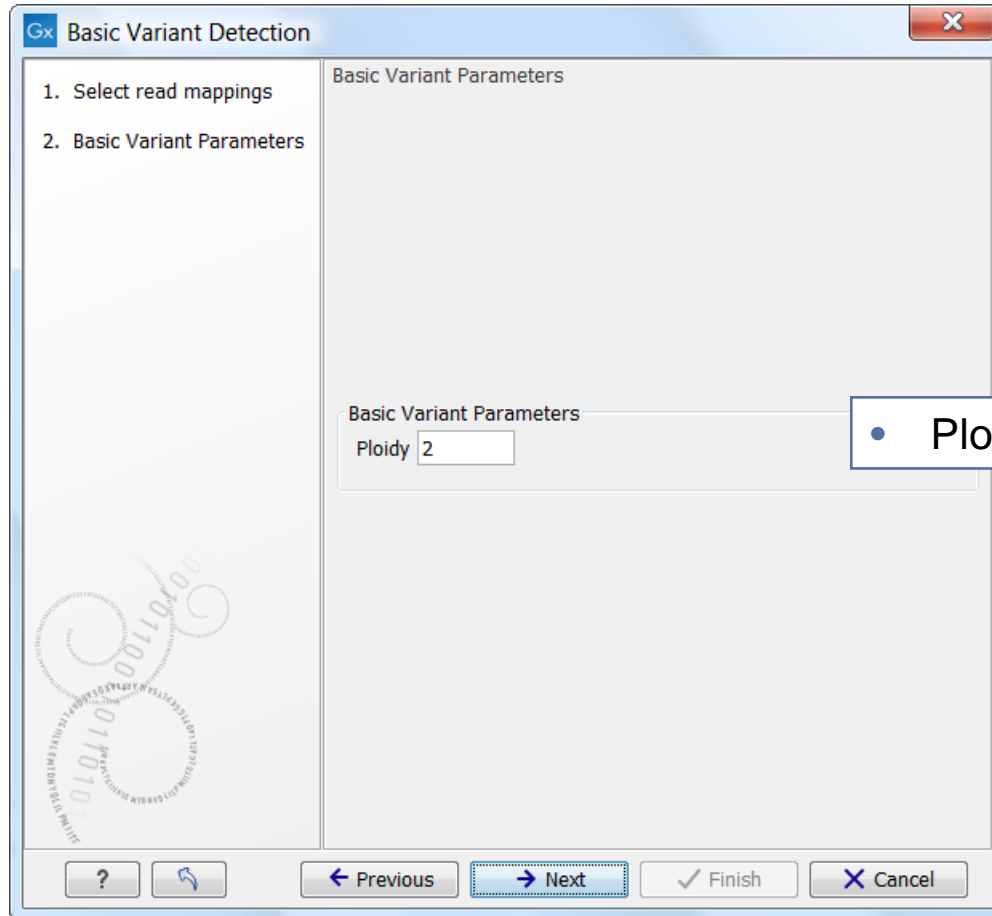
In homopolymer regions with minimum length

With frequency below

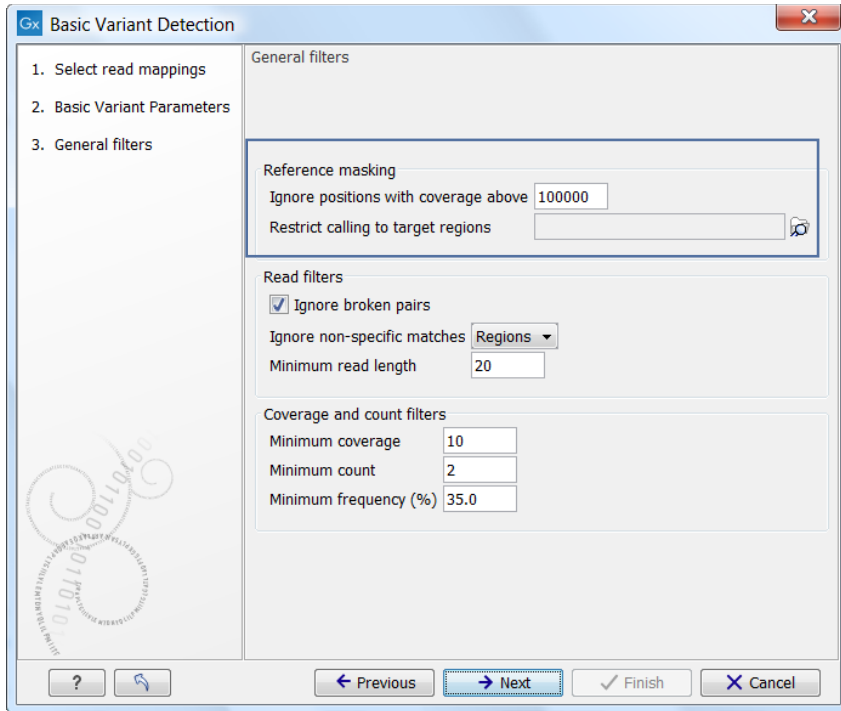
?
↶
↷ Next
✓ Finish
✗ Cancel



- 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。

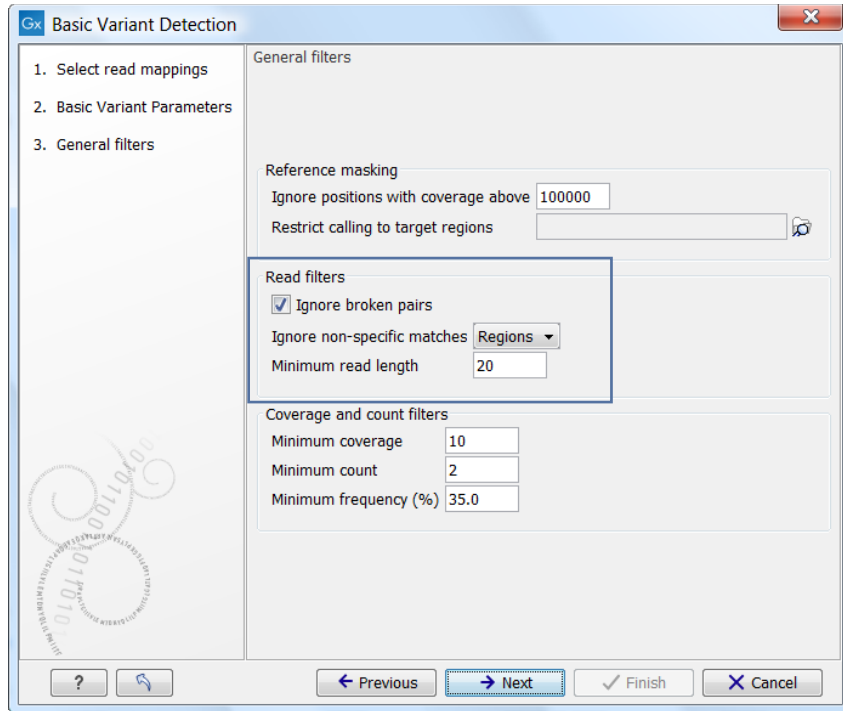


- Ploidy: 参照配列の倍数性



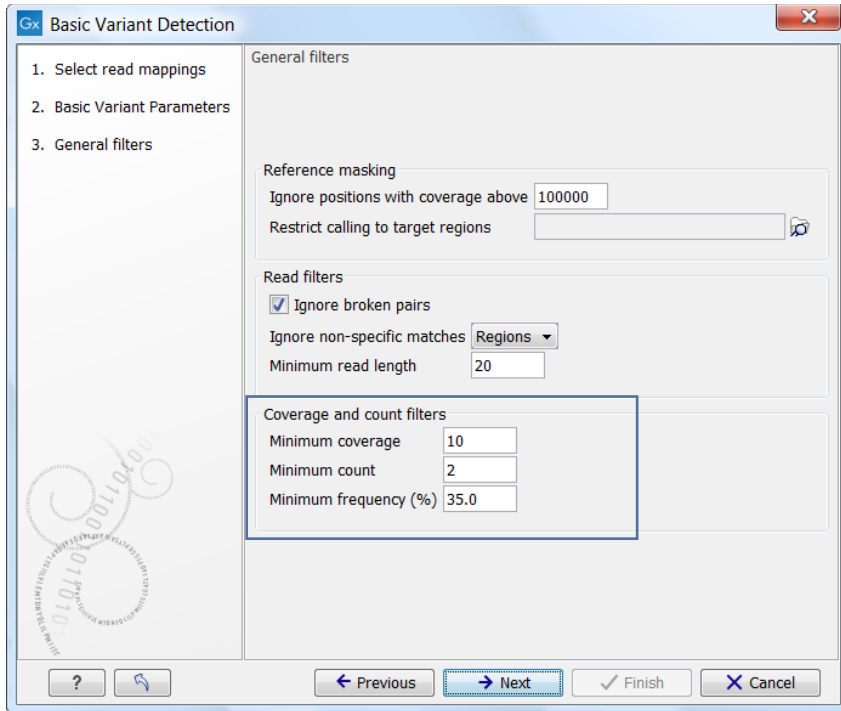
Reference masking

- Ignore positions with coverage above : カバレッジが指定した数字以上のバリエーションについてリストに含めない
- Restrict calling to target regions : バリエーション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)



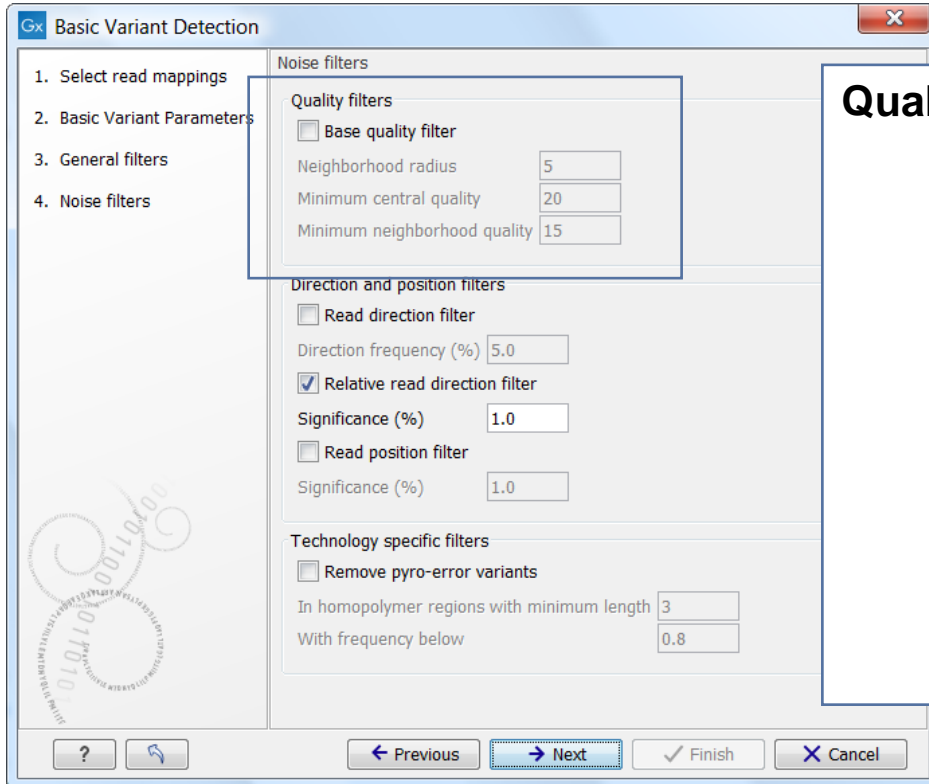
Read filters

- 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エーション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、その領域のバリエーションを検出しません。
- 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、その長さからnon-specificになる可能性があるためです。



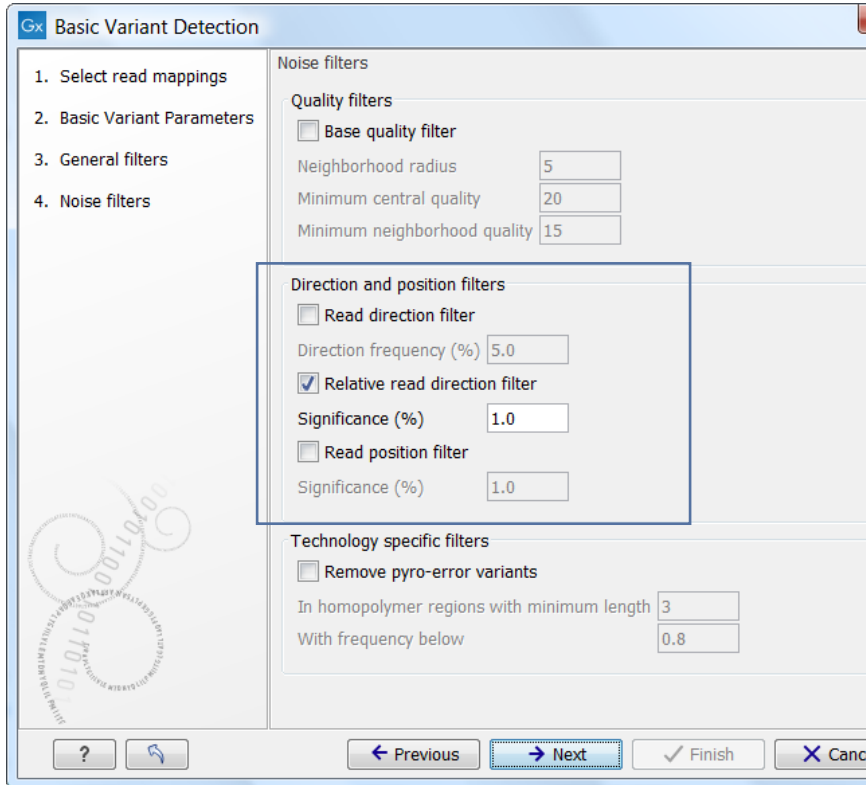
Coverage and count filters

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



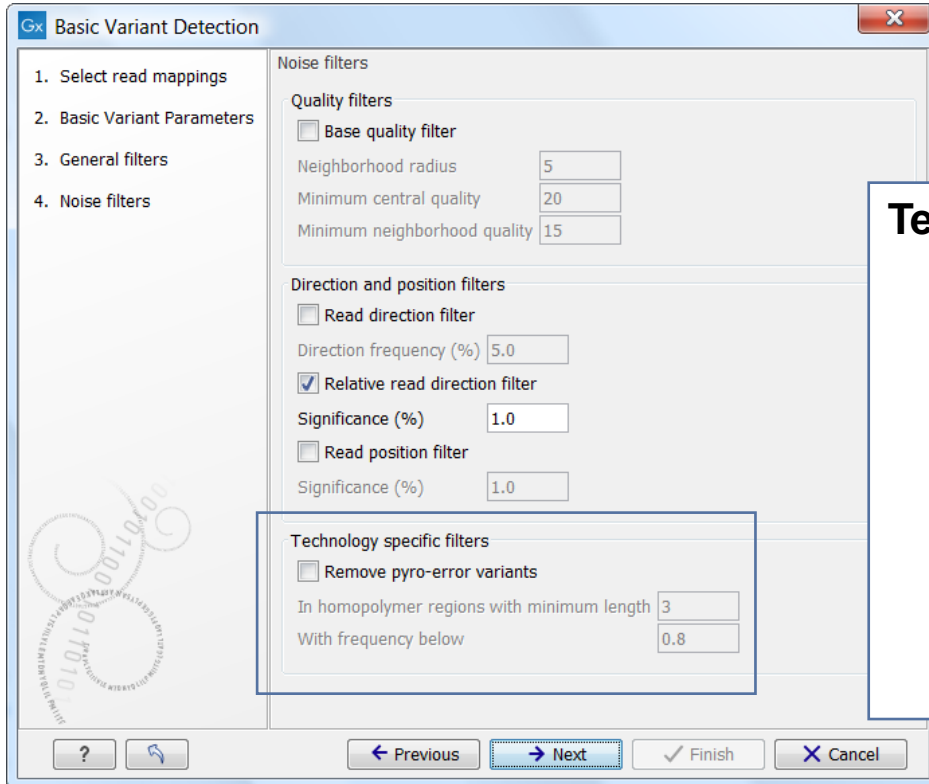
Quality filter

- 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)



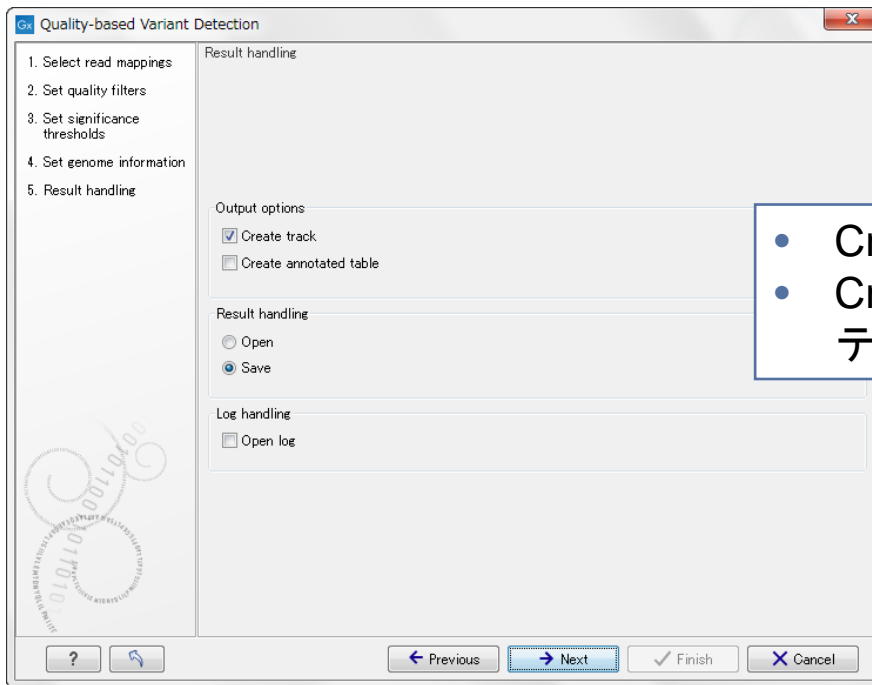
Direction and position filters :

- 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 : システム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、バリエーション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

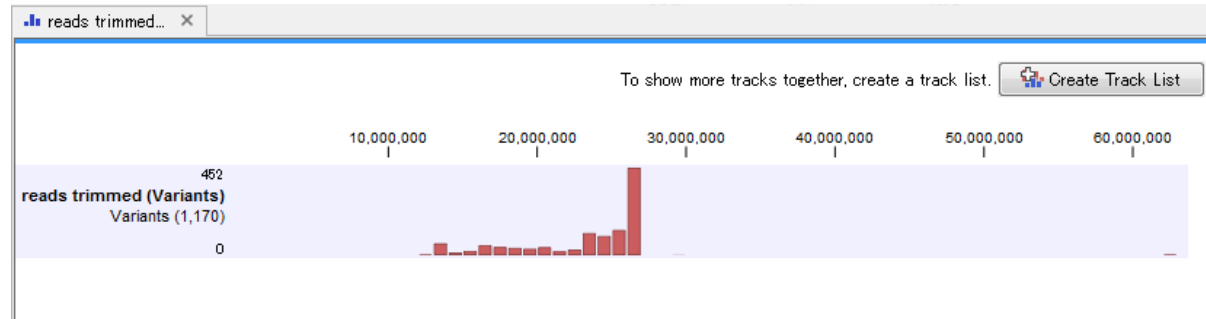
- Remove pyro-error variants : ホモポリマー領域に対するエラーの除去
 - In homopolymer regions with minimum length : 指定した長さのホモポリマー領域のInDelを取り除く。
 - With frequency below : 指定した頻度以下のものについてのみフィルターを適用。



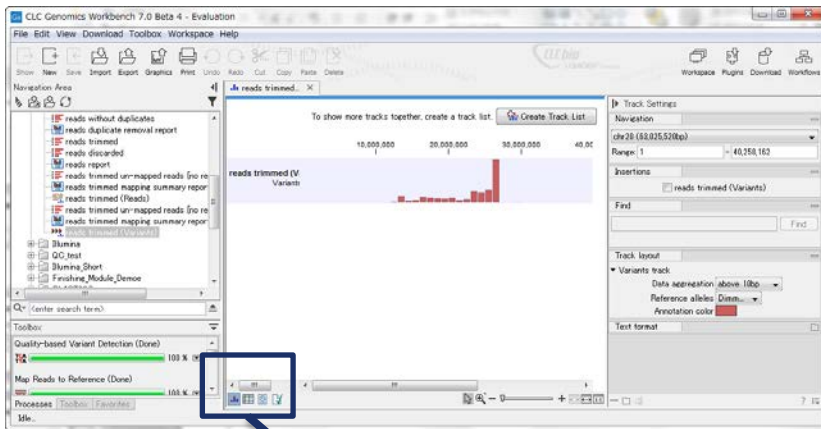
- Create track: トラックの作成
- Create annotated table: アノテーション付のテーブルの作成

結果

reads trimmed (Variants)



- 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。



Chromos.	Region	Type	Refers...	Allele	Refers...	Length	Zeposity	Count	Covers...	Freque...	Forwar...	Revers...	Forwar...	Average	Hyper...
chr20	12668235	SNV	T	C	No	1	Homozyg...	2	2	100.00	1	1	0.50	25.00	No
chr20	12681535	SNV	G	A	No	1	Homozyg...	2	2	100.00	1	1	0.50	20.00	No
chr20	12718797	SNV	G	A	No	1	Homozyg...	2	2	100.00	1	1	0.50	26.00	No
chr20	12771939	SNV	T	C	No	1	Homozyg...	2	2	100.00	1	1	0.50	26.00	No
chr20	12818551	SNV	C	G	No	1	Homozyg...	2	2	100.00	1	1	0.50	32.00	No
chr20	13029784	SNV	A	C	No	1	Homozyg...	72	72	100.00	30	39	0.45	24.70	No
chr20	13029790	SNV	C	T	No	1	Heterozyg...	40	70	57.14	18	22	0.45	24.20	No
chr20	13029799	SNV	C	T	No	1	Heterozyg...	30	70	42.86	18	12	0.40	23.80	No
chr20	13029802	SNV	C	T	No	1	Homozyg...	57	57	100.00	16	41	0.28	24.20	No
chr20	13029920	SNV	C	G	No	1	Homozyg...	40	40	100.00	24	16	0.40	24.70	No
chr20	13055818	SNV	T	G	No	1	Homozyg...	72	72	100.00	36	36	0.50	29.51	No
chr20	13058035	SNV	A	G	No	1	Homozyg...	59	89	100.00	30	39	0.42	29.74	No
chr20	13071871	SNV	C	A	No	1	Heterozyg...	24	52	54.04	2	22	0.08	30.85	No
chr20	13071871	SNV	C	C	Yes	1	Heterozyg...	28	82	45.16	3	25	0.11	31.25	No
chr20	13074235	SNV	G	A	No	1	Homozyg...	45	45	100.00	0	37	0.10	30.67	No
chr20	13082478	SNV	C	C	Yes	1	Heterozyg...	5	12	41.67	4	1	0.20	25.80	No
chr20	13082478	SNV	G	T	No	1	Heterozyg...	7	12	58.33	5	2	0.29	24.47	No
chr20	13082493	SNV	T	C	No	1	Homozyg...	10	10	100.00	3	7	0.30	27.70	No
chr20	13082892	SNV	C	T	No	1	Homozyg...	2	2	100.00	1	1	0.50	22.00	No
chr20	13145538	SNV	T	C	No	1	Homozyg...	2	2	100.00	1	2	0.33	31.00	No
chr20	13166599	SNV	G	A	No	1	Homozyg...	2	2	100.00	1	1	0.50	20.00	No

- 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。

Chrom...	Region	Type	Ref...	Allele	Ref...	Len...	Zygoty...	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

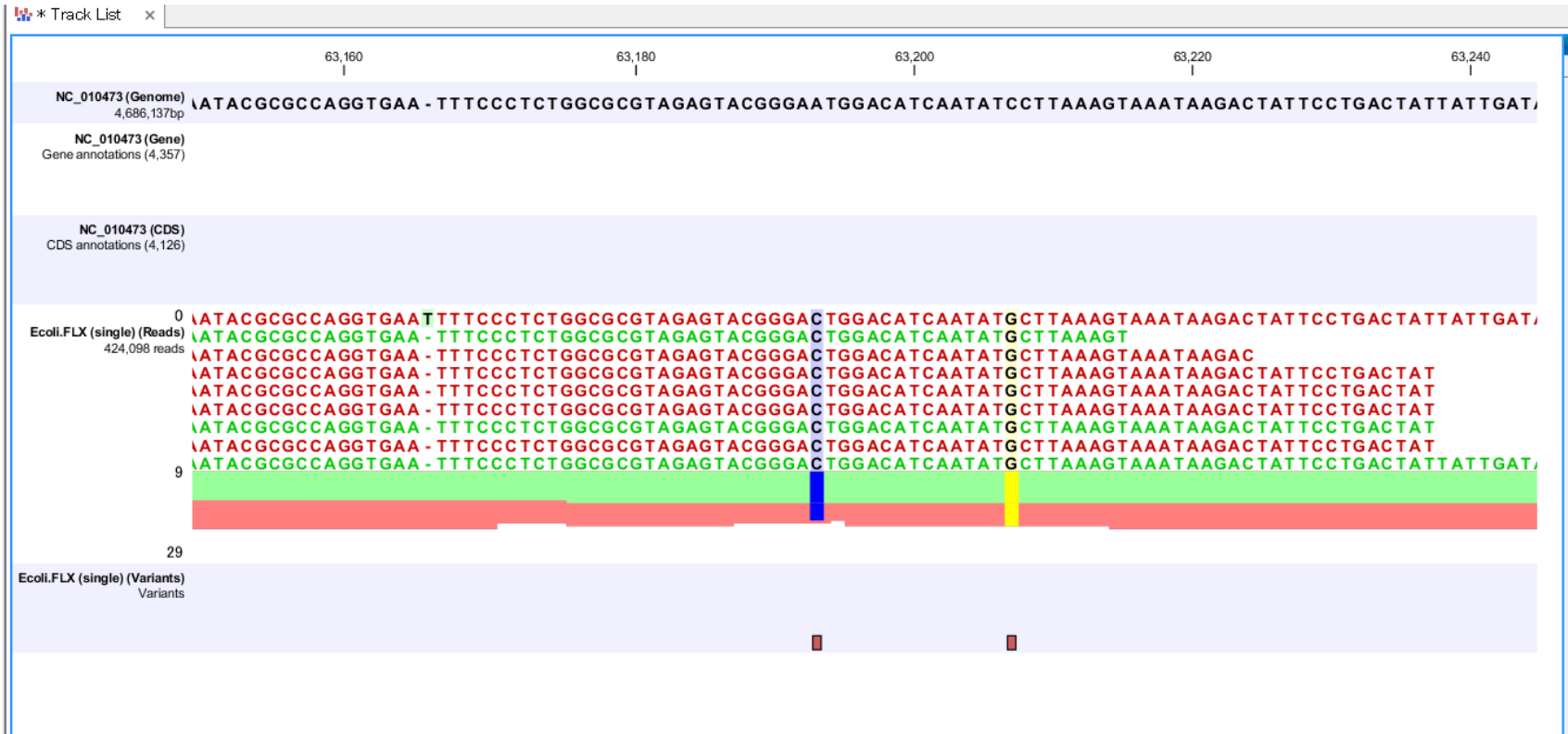
- Count: クオリティのフィルターをパスしたリードの数
- Coverage: クオリティのフィルターをパスしたリードの数
- Frequency: バリエントが見られた頻度
- Probability: バリエントのアレルの事後確率（そのアレルが尤もであるとする確率。高い方がより確度が高いという事。）
- Forward reads: その領域に見られたForwardリードの数
- Reverse reads: その領域に見られたReverseリードの数
- Forward/reverse: Forward/Total reads または Reverse/Total reads のうち小さい方の値。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: バリエ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: バリエ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リエ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エントの塩基に差がある。
- Hyper-allelic : 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

Show column

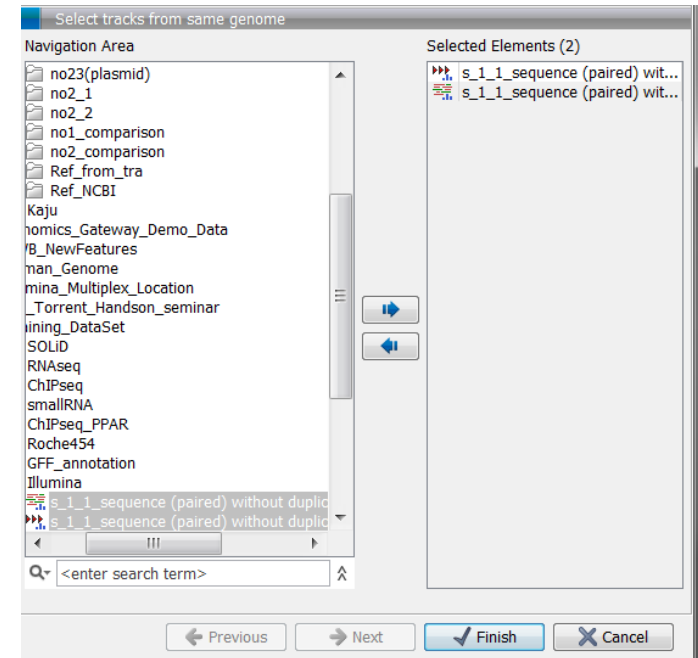
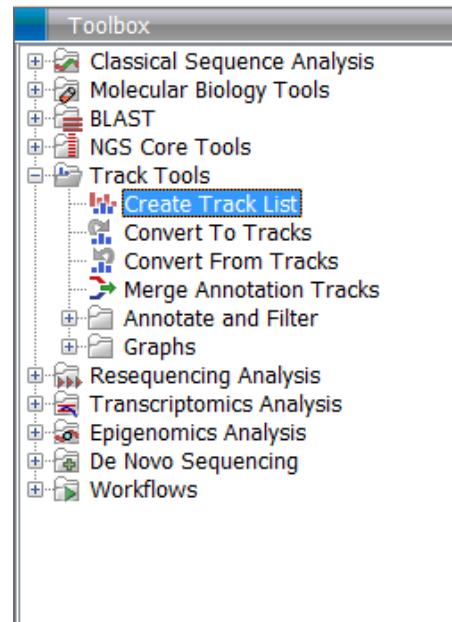
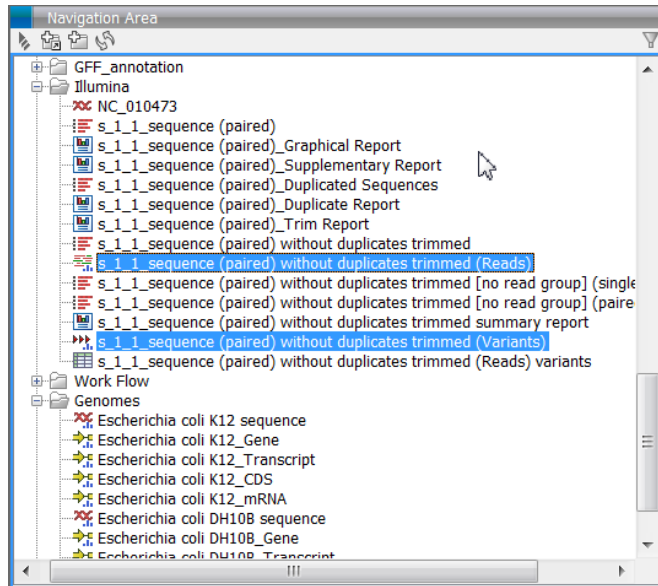
- Chromosome
- Region
- Type
- Reference
- Allele
- Reference allele
- Length
- Linkage
- Zygoty...
- Count
- Coverage
- Frequency
- Probability
- Forward read count
- Reverse read count
- Forward/reverse balance
- Average quality
- Read count
- Read coverage
- # unique start positions
- # unique end positions
- BaseQRankSum
- Read position test probability
- Read direction test probability
- Hyper-allelic
- Homopolymer

Select All

トラックリストの作成

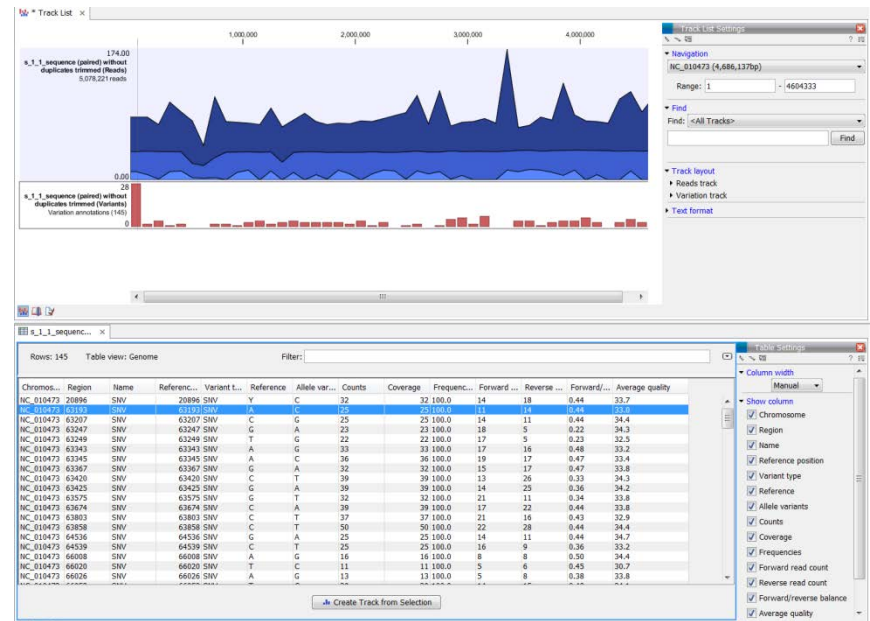
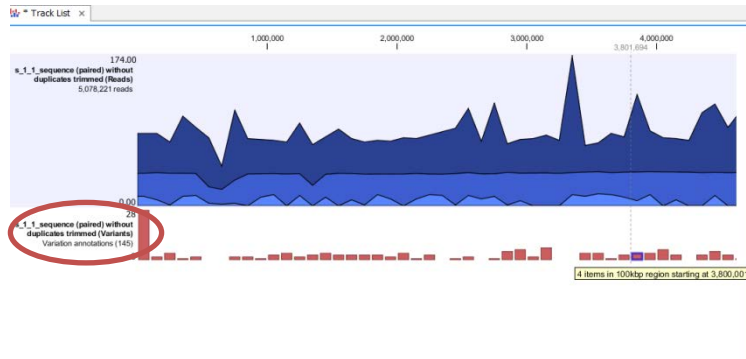


トラックリスト作成



- Navigation Areaからマッピングデータとバリエーションの結果を選択。
- 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) を使ったバリエーション検出

Reference



A

?

A

A

T

T

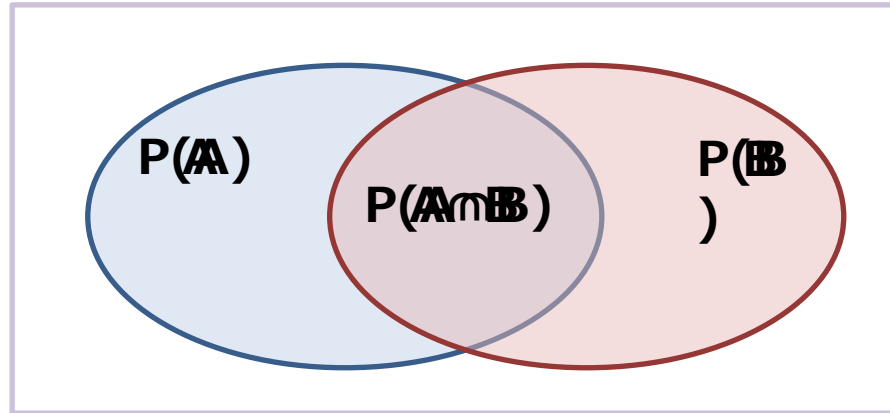
C

?: Site type (ex) A/A, A/T, A/C
... ?



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、バリエーションとして結果返す。



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

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

$$P(B | A)P(A) = P(A | B)P(B)$$

ベイズの定理

事後確率
Posterior

←

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

←

事前確率
Prior

←

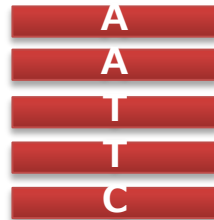
尤度
Likelihood

Reference



A

?



?: Site type (ex) A/A, A/T, A/C
... ?

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

S : Site type

R : Reads

$P(R | S)$: **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。

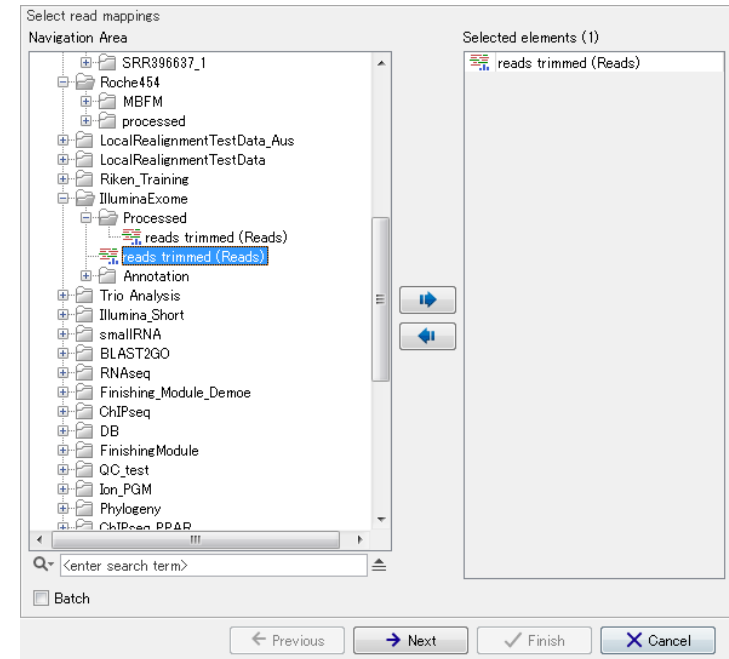
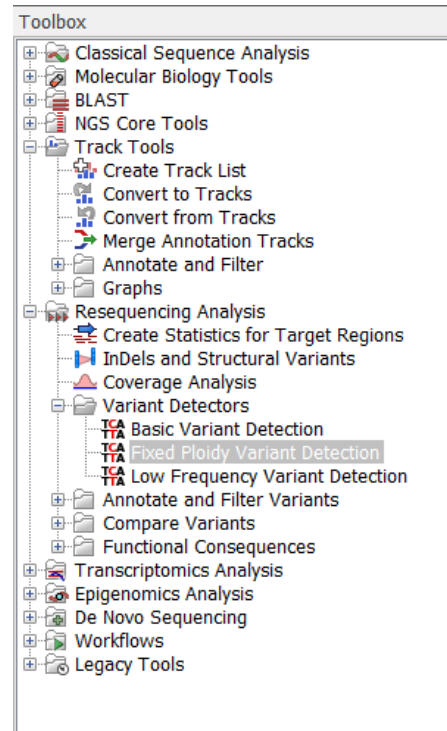
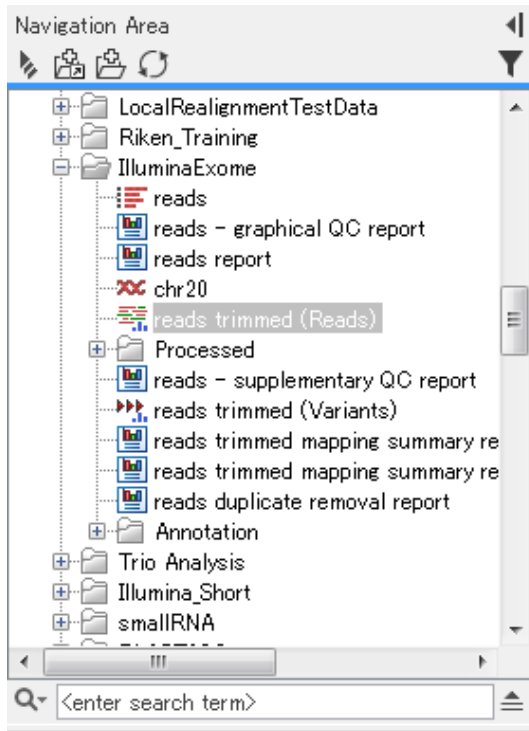
EMアルゴリズム (Expectation Maximization algorithm) は、得られたデータから推定したい現象が観察できない場合に、その確率を推定する、一般的な統計の手法。

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

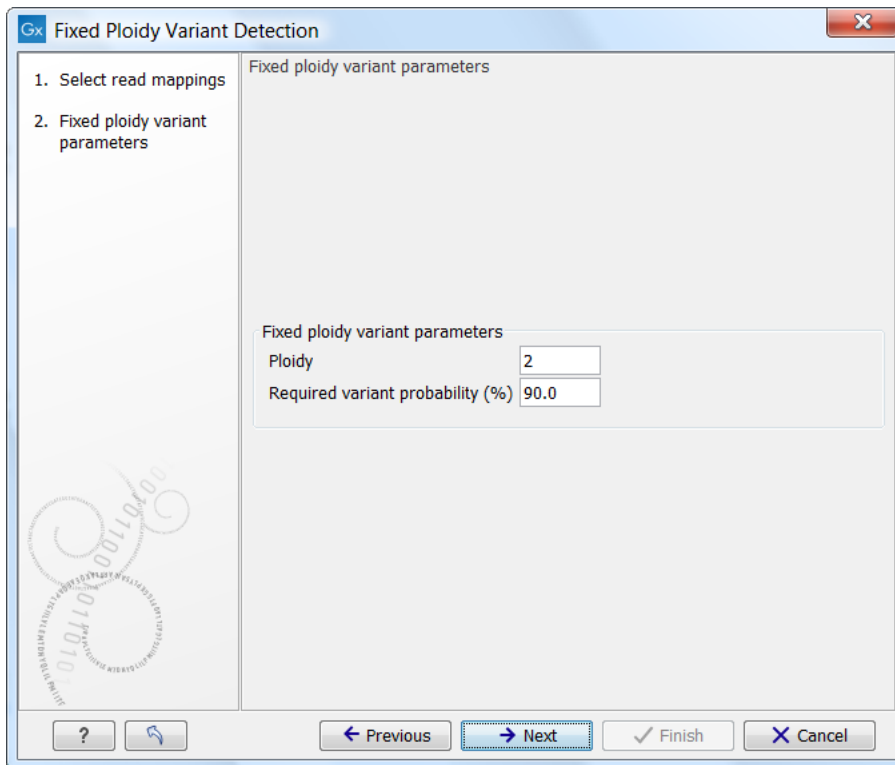
Error Model

- 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
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



- 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


1. Select read mappings

2. Fixed ploidy variant parameters

Fixed ploidy variant parameters

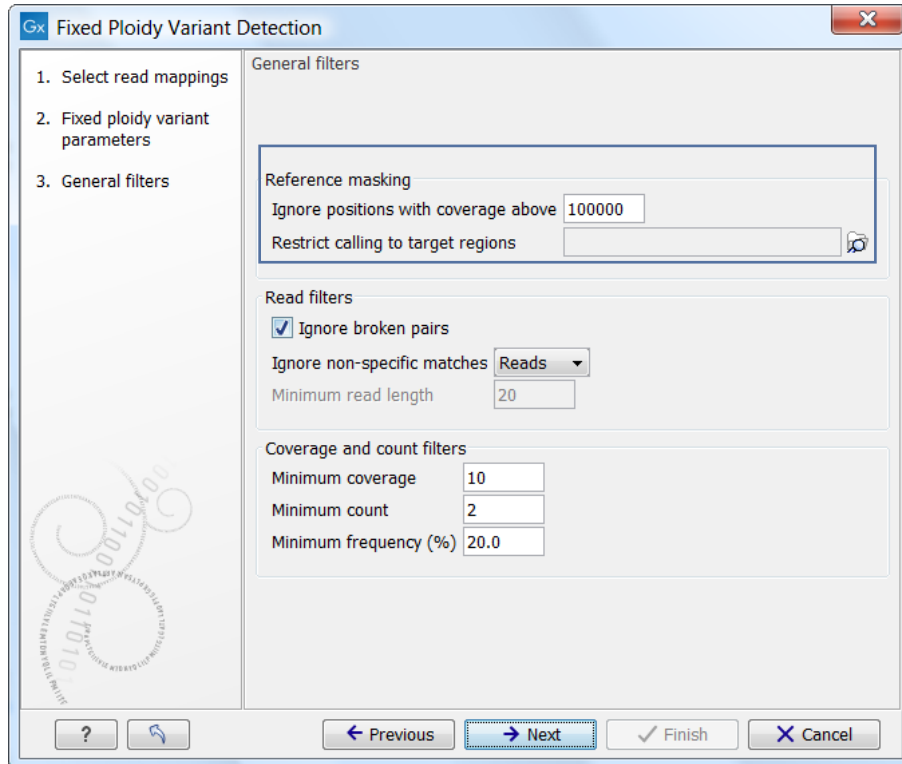
Ploidy

Required variant probability (%)

? 

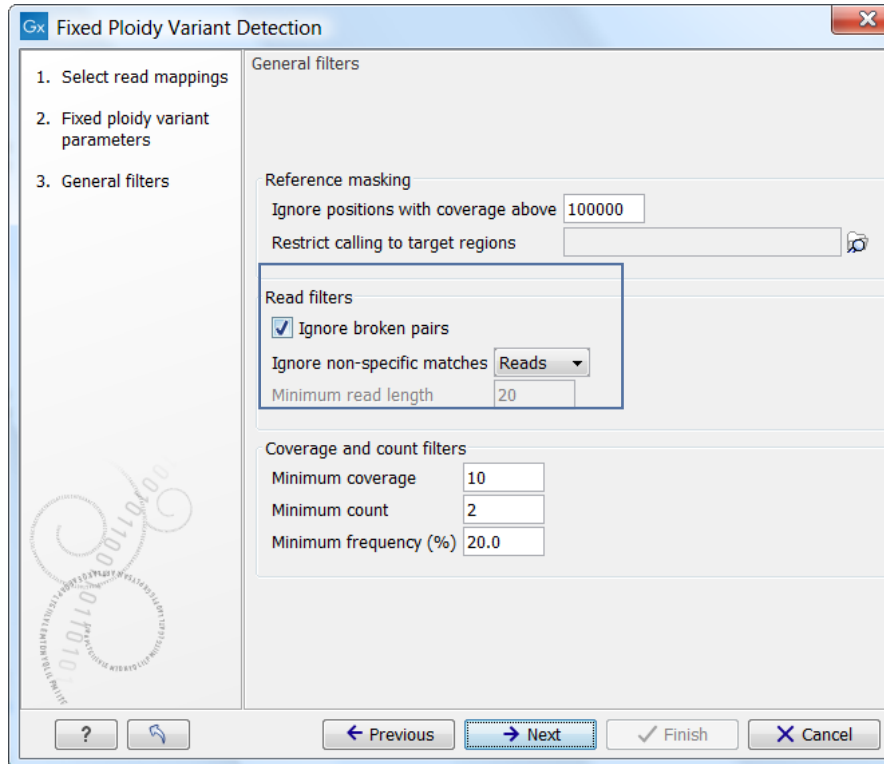
Ploidy : 参照配列の倍数性

- Required variant probability : バリエ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、検出されるバリエ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エーションについてリストに含めない
- Restrict calling to target regions : バリエーション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エーション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、その領域のバリエーションを検出しません。
- 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、その長さからnon-specificになる可能性があるためです。

Fixed Ploidy Variant Detection

- Select read mappings
- Fixed ploidy variant parameters
- General filters

General filters

Reference masking

Ignore positions with coverage above

Restrict calling to target regions

Read filters


Ignore broken pairs

Ignore non-specific matches

Minimum read length

Coverage and count filters

Minimum coverage	<input type="text" value="10"/>
Minimum count	<input type="text" value="2"/>
Minimum frequency (%)	<input type="text" value="20.0"/>

? 

Coverage and count filters

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

Gx Fixed Ploidy Variant Detection

- Select read mappings
- Fixed ploidy variant parameters
- General filters
- Noise filters

Noise filters

Quality filters

Base quality filter

Neighborhood radius

Minimum central quality

Minimum neighborhood quality

Direction and position filters

Read direction filter

Direction frequency (%)

Relative read direction filter

Significance (%)

Read position filter

Significance (%)

Technology specific filters

Remove pyro-error variants

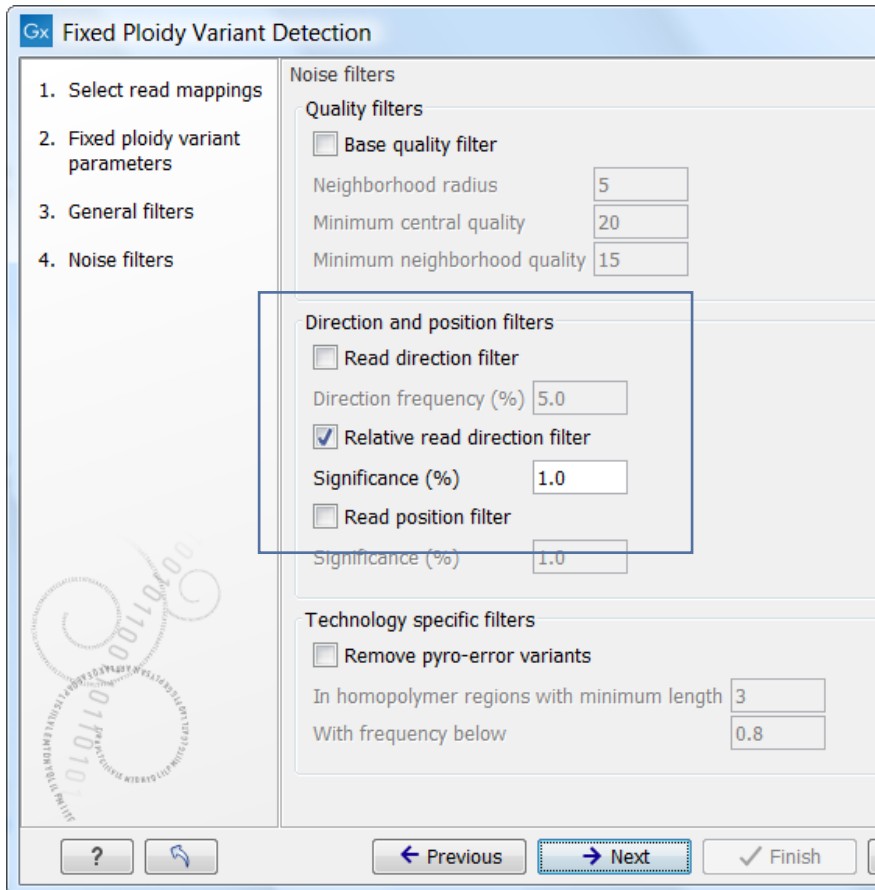
In homopolymer regions with minimum length

With frequency below

? ↶ ↷ ✓ Finish ✕ Cancel

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 : システム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、バリエーション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。

Gx Fixed Ploidy Variant Detection
X

1. Select read mappings
2. Fixed ploidy variant parameters
3. General filters
4. Noise filters

Noise filters

Quality filters

Base quality filter

Neighborhood radius

Minimum central quality

Minimum neighborhood quality

Direction and position filters

Read direction filter

Direction frequency (%)

Relative read direction filter

Significance (%)

Read position filter

Significance (%)

Technology specific filters

Remove pyro-error variants

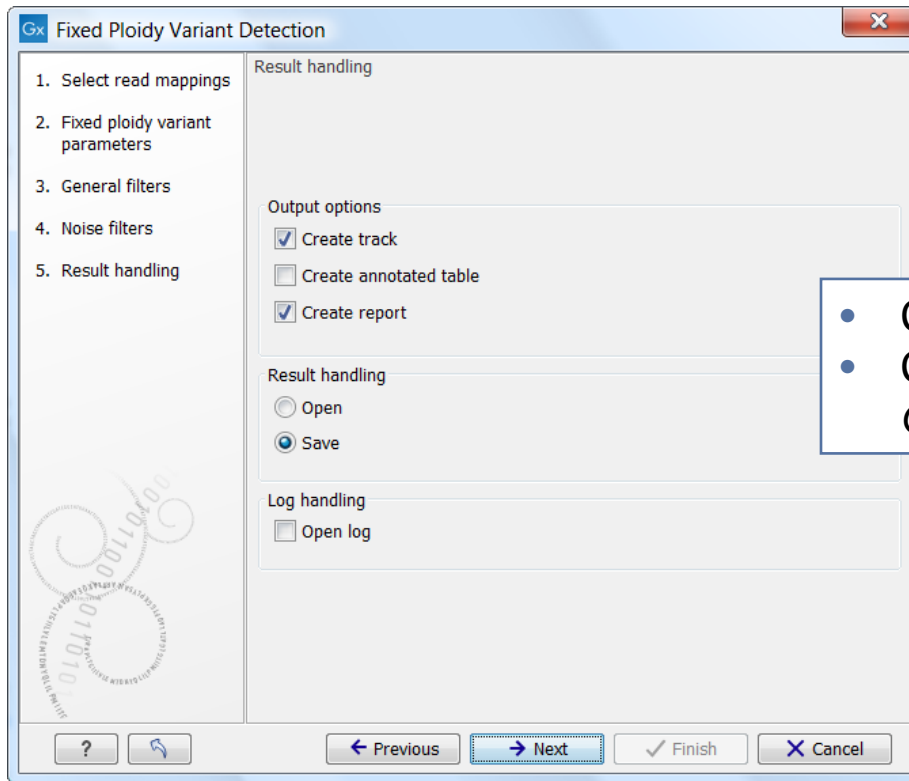
In homopolymer regions with minimum length

With frequency below

?
↶ Previous
Next
✓ Finish
X Cancel

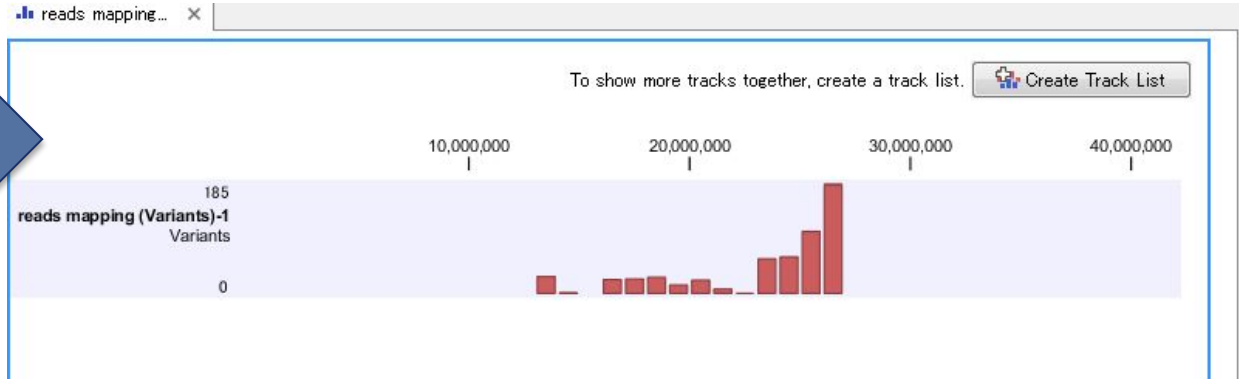
Technology specific filters

- 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

reads mapping (Variants)-



reads mapping... x

Rows: 590 Table view: Genome Filter:

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	58.94	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	29.56	
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	29.79	
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	

Table Settings

Column width: Manual

Show column:

- Chromosome
- Region
- Type
- Reference
- Allele
- Reference allele
- Linkage
- Zygosity
- Count
- Coverage
- Frequency
- Probability
- Forward read count
- Reverse read count
- Forward/reverse balance
- Average quality

Select All
Deselect All



Chrom...	Region	Type	Ref...	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

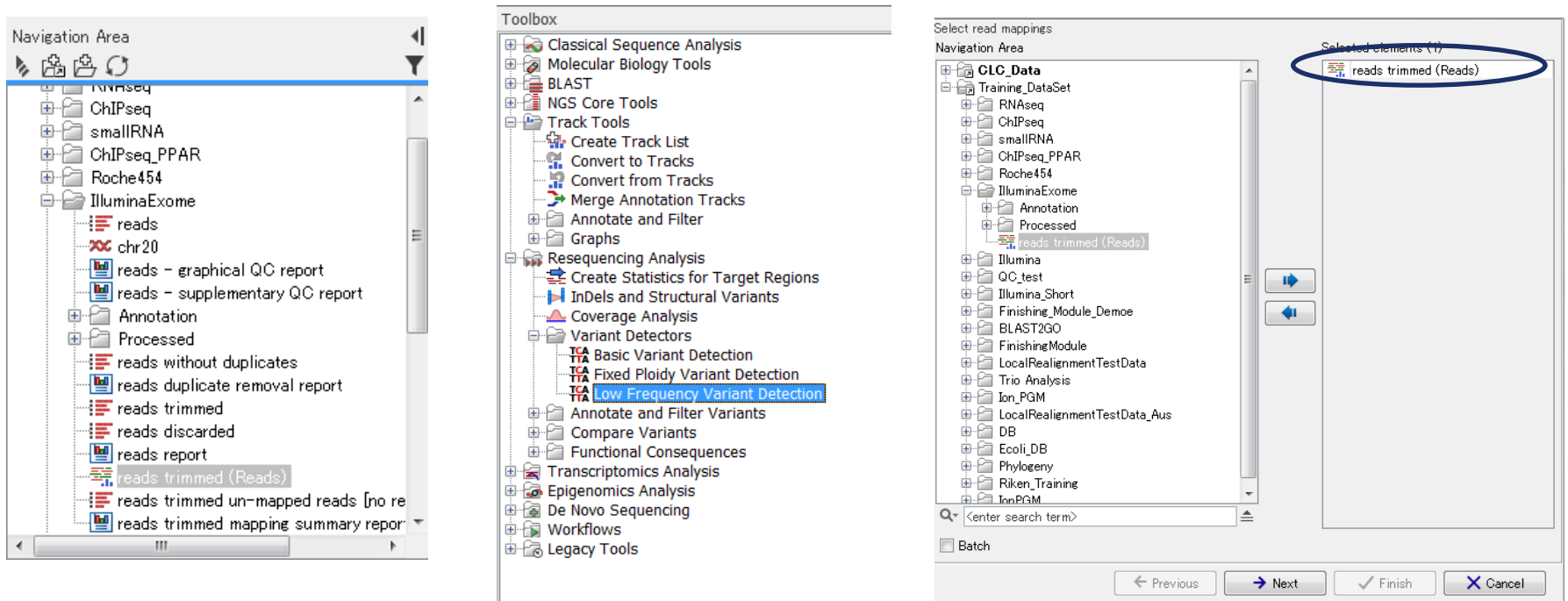
- Count: クオリティのフィルターをパスしたリードの数
- Coverage: クオリティのフィルターをパスしたリードの数
- Frequency: バリエントが見られた頻度
- Probability: バリエントのアレルの事後確率（そのアレルが尤もであるとする確率。高い方がより確度が高いという事。）
- Forward reads: その領域に見られたForwardリードの数
- Reverse reads: その領域に見られたReverseリードの数
- Forward/reverse: Forward/Total reads または Reverse/Total reads のうち小さい方の値。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: バリエ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: バリエ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リエ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エントの塩基に差がある。
- Hyper-allelic : 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

Show column

- Chromosome
- Region
- Type
- Reference
- Allele
- Reference allele
- Length
- Linkage
- Zygosity
- Count
- Coverage
- Frequency
- Probability
- Forward read count
- Reverse read count
- Forward/reverse balance
- Average quality
- Read count
- Read coverage
- # unique start positions
- # unique end positions
- BaseQRankSum
- Read position test probability
- Read direction test probability
- Hyper-allelic
- Homopolymer

Select All

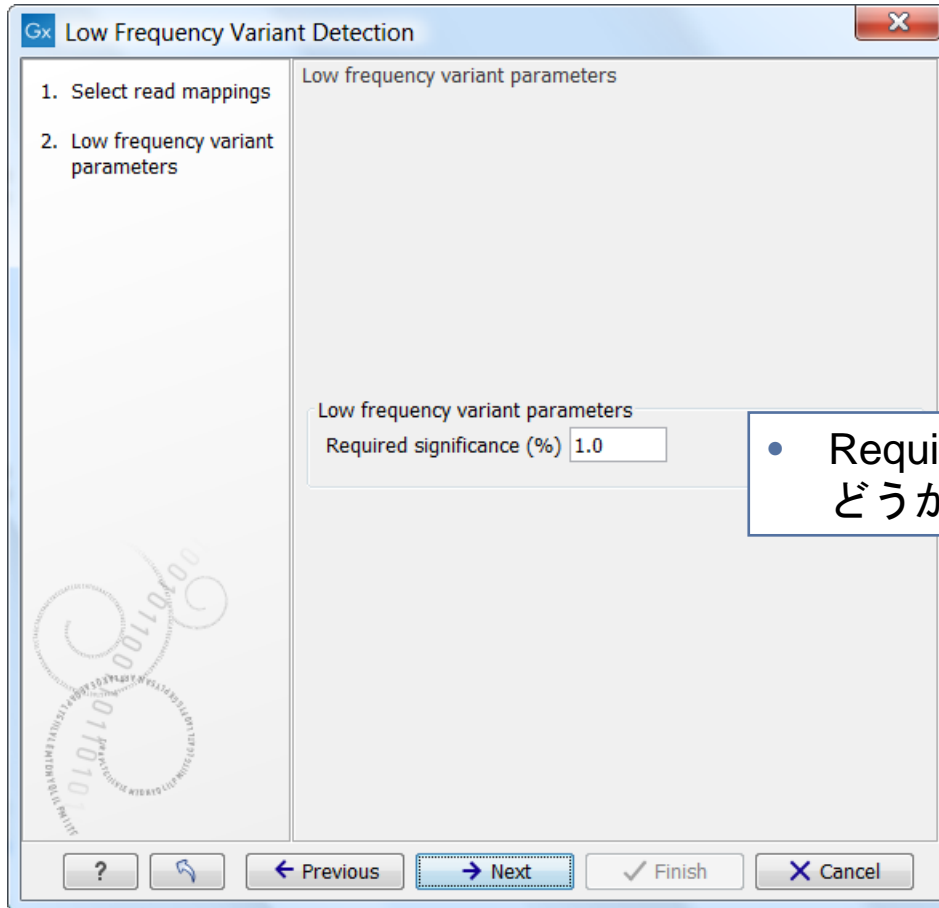
- Low frequency Variant Detection では、倍数性を仮定せず、対象となる領域が、シーケンスエラーなのか、そうではない (=バリエーション) なのかを検定しています。
- Error モデルについては、Fixed Ploidy Variant Detection にて採用したエラーモデルを使い、計算し、尤度比検定を行っています。



The screenshot shows the software interface with the following components:

- Navigation Area (Left):** A tree view showing a project structure. Under 'IlluminaExome', the 'reads trimmed (Reads)' folder is selected and highlighted in blue.
- Toolbox (Middle):** A list of analysis tools. Under 'Resequencing Analysis' > 'Variant Detectors', 'Low Frequency Variant Detection' is selected and highlighted in blue.
- Select read mappings dialog (Right):** A dialog box titled 'Select read mappings' with a 'Navigation Area' on the left and a 'Selected elements (1)' list on the right. The 'reads trimmed (Reads)' element is selected in the list and circled in red. The dialog also features a search bar, a 'Batch' checkbox, and navigation buttons: 'Previous', 'Next', 'Finish', and 'Cancel'.

- 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。



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

Fixed Ploidy Variant Detection


1. Select read mappings
2. Fixed ploidy variant parameters
3. General filters

General filters

Reference masking
Ignore positions with coverage above
Restrict calling to target regions

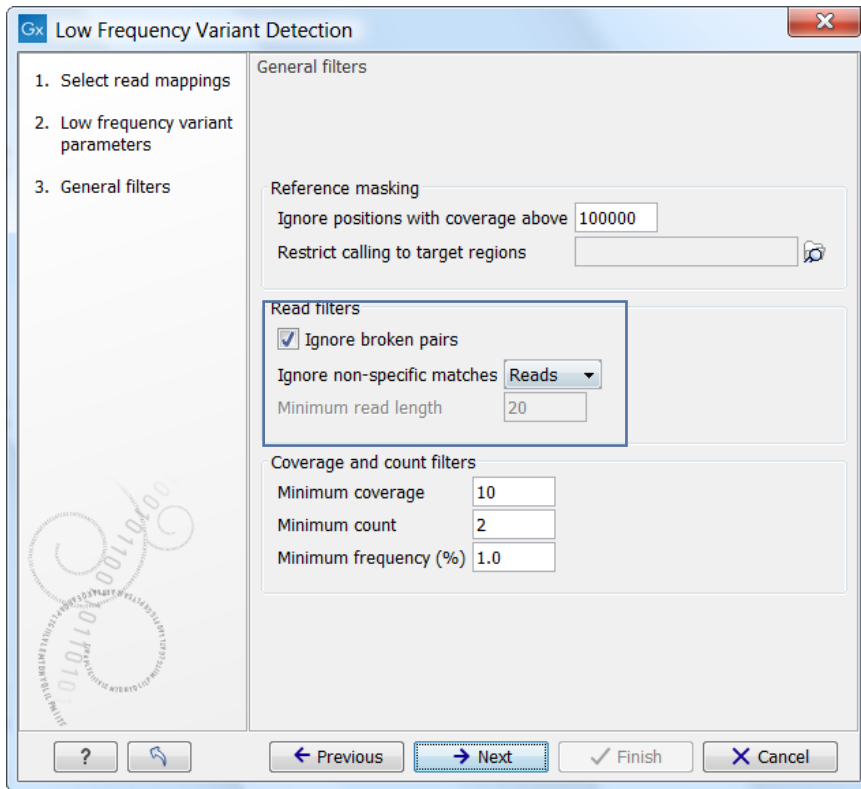
Read filters
 Ignore broken pairs
Ignore non-specific matches
Minimum read length

Coverage and count filters
Minimum coverage
Minimum count
Minimum frequency (%)

? 

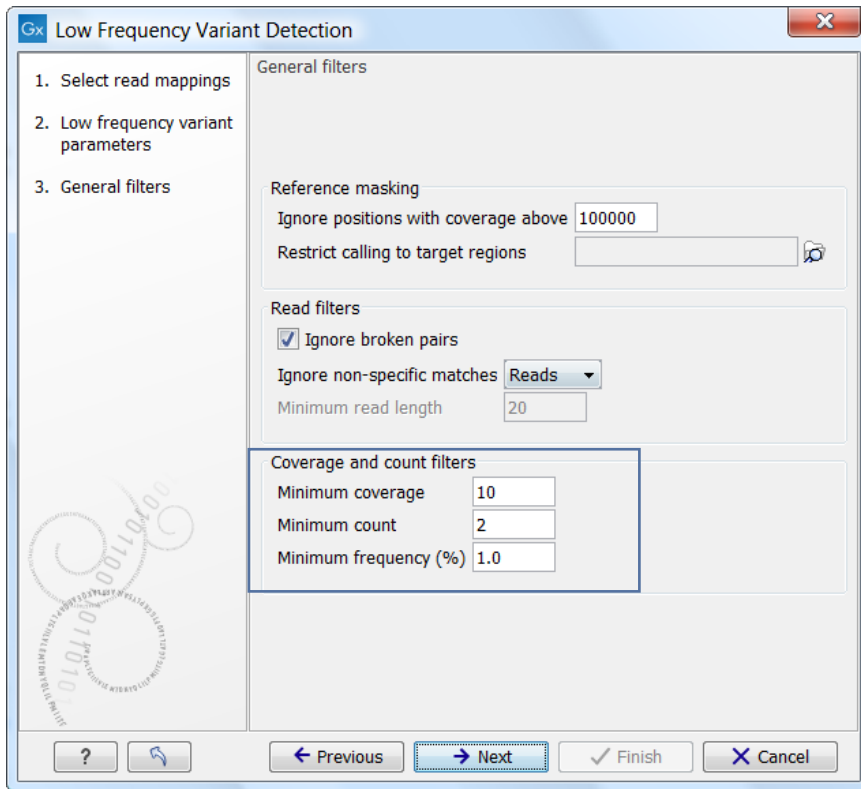
Reference masking

- Ignore positions with coverage above : カバレッジが指定した数字以上のバリエーションについてリストに含めない
- Restrict calling to target regions : バリエーション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)



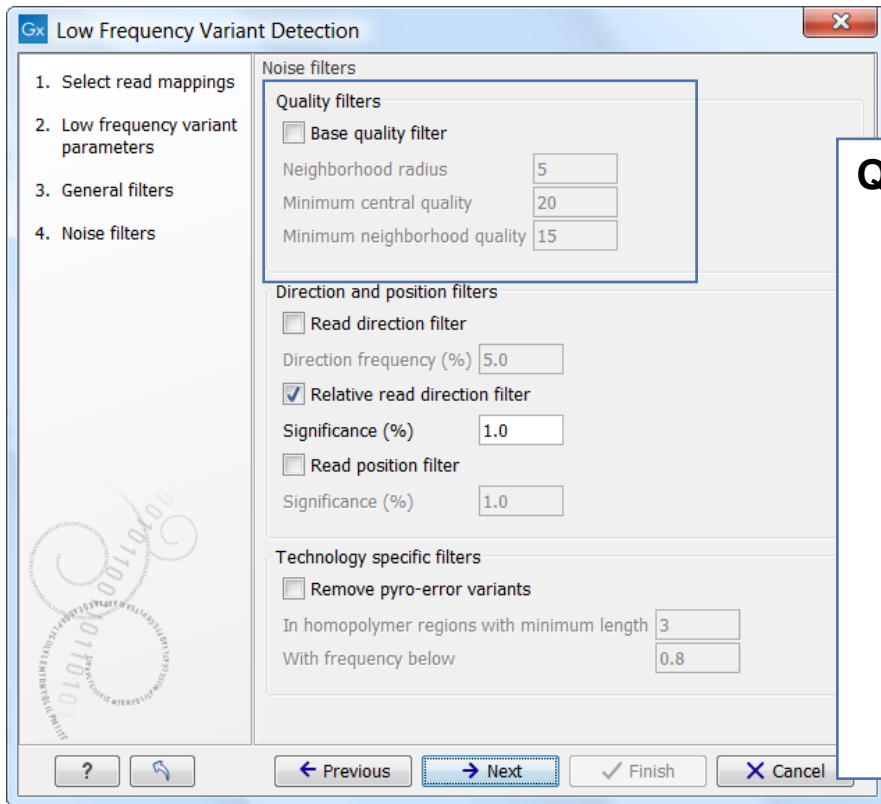
Read filters

- 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エーション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、その領域のバリエーションを検出しません。
- 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、その長さからnon-specificになる可能性があるためです。



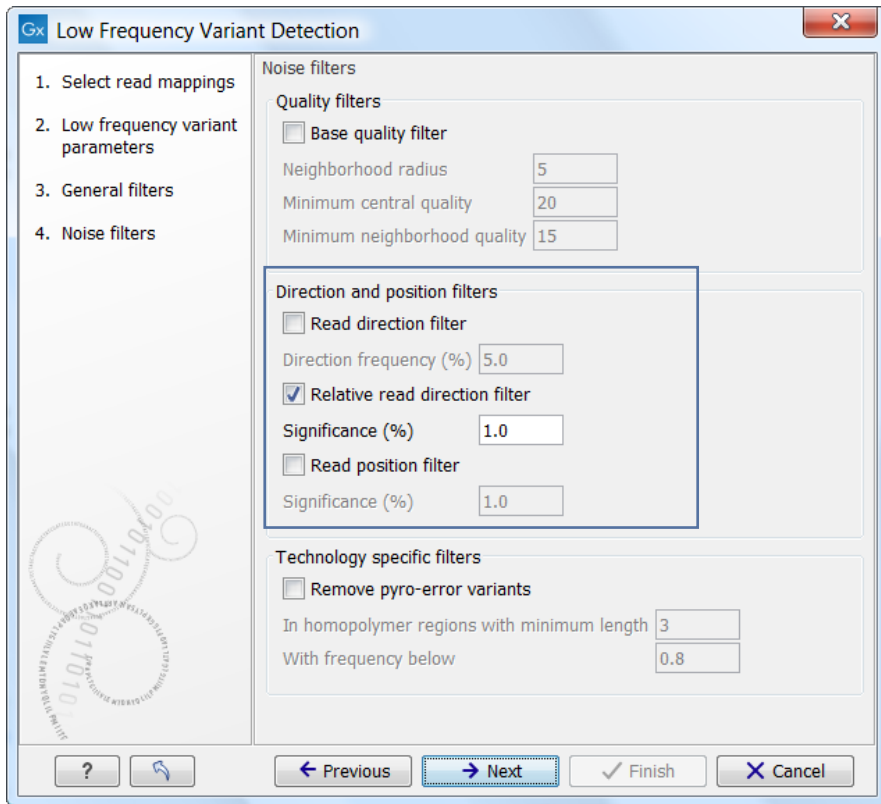
Coverage and count filters

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



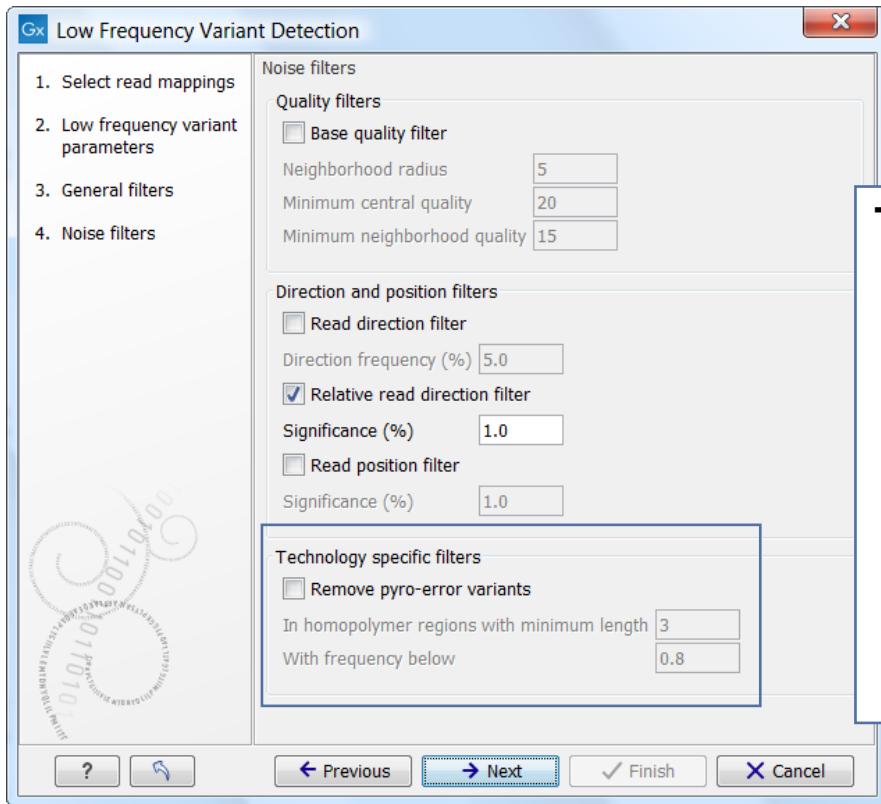
Quality filter

- 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)



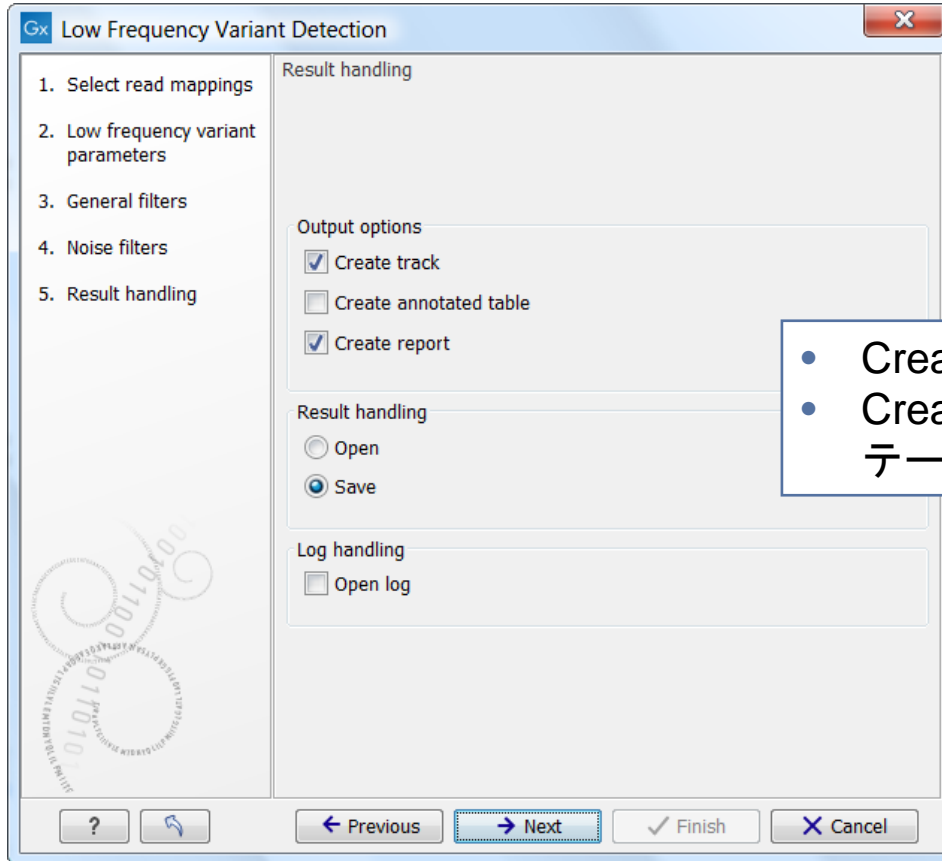
Direction and position filters :

- 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 : システム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、バリエーション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。



- 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

Rows: 231 Table view: Genome

Chrom...	Region	Type	Ref...	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

Show column

- Chromosome
- Region
- Type
- Reference
- Allele
- Reference allele
- Length
- Linkage
- Zygosity
- Count
- Coverage
- Frequency
- Probability
- Forward read count
- Reverse read count
- Forward/reverse balance
- Average quality
- Read count
- Read coverage
- # unique start positions
- # unique end positions
- BaseQRankSum
- Read position test probability
- Read direction test probability
- Hyper-allelic
- Homopolymer

Select All

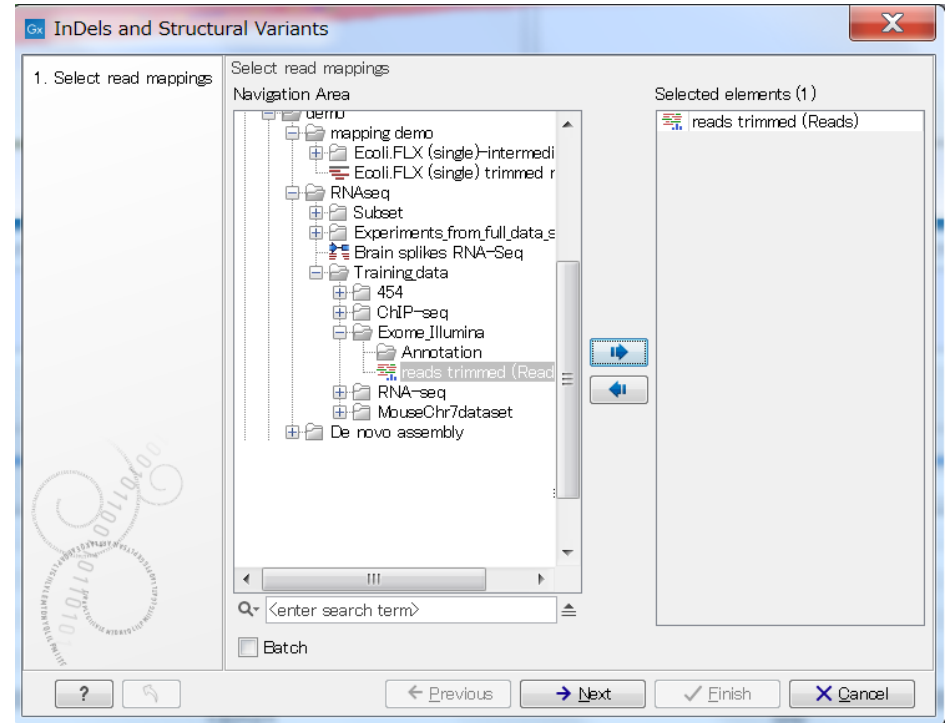
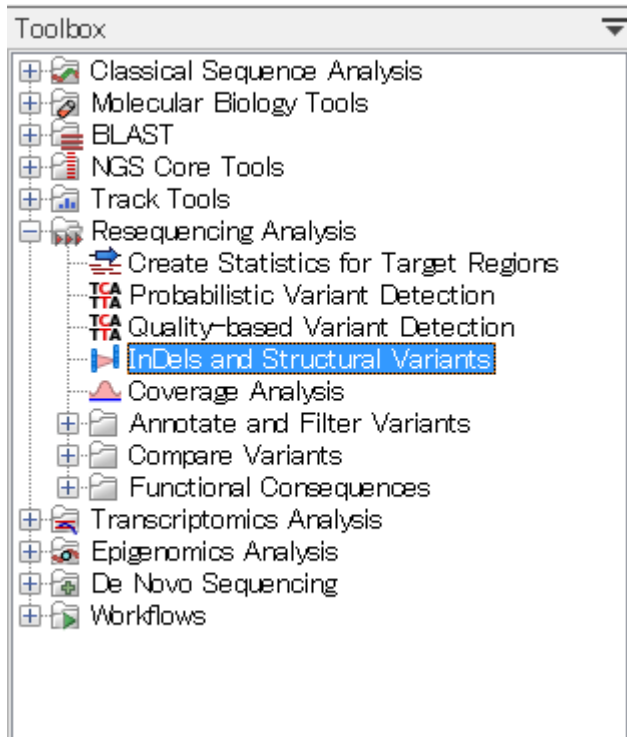
- Count: クオリティのフィルターをパスしたリードの数
- Coverage: クオリティのフィルターをパスしたリードの数
- Frequency: バリエントが見られた頻度
- Probability: バリエントのアレルの事後確率（そのアレルが尤もであるとする確率。高い方がより確度が高いという事。）
- Forward reads: その領域に見られたForwardリードの数
- Reverse reads: その領域に見られたReverseリードの数
- Forward/reverse: Forward/Total reads または Reverse/Total reads のうち小さい方の値。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: バリエ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: バリエ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リエ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エントの塩基に差がある。
- Hyper-allelic : 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

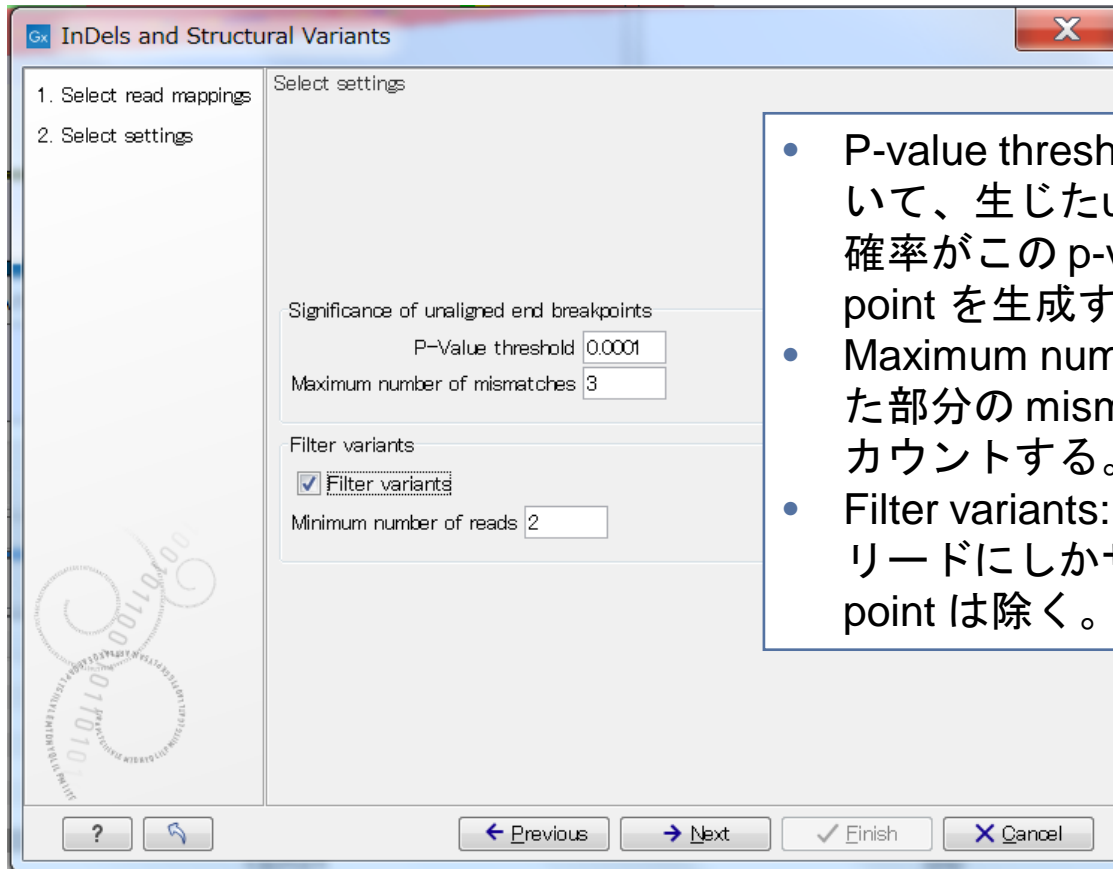
- Quality Based Variant Detection や Probabilistic Variant Detection では変異やInDelを検出できませんでした。
- しかしながら大きなInDelの検出や構造変異については、上記ツールでの検出は難しい場合があります。



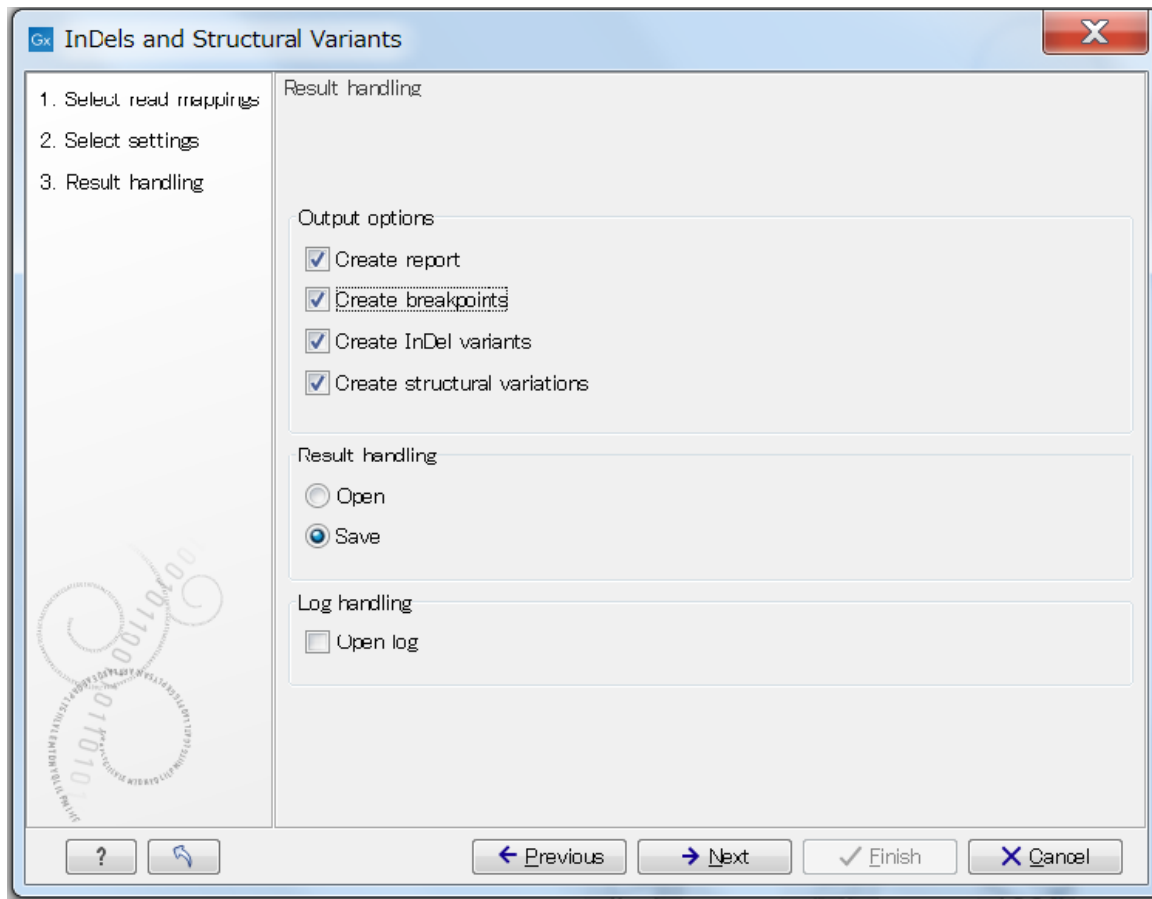
- アルゴリズムにとっては、大きなInsertionやDeletionを受け入れるよりは、Unaligned end とするほうがスコアを大きくできるからです。
- InDel and Structural Variants ツールでは、このUnaligned end に着目して、大きなInDelや構造変異を見つけます。
- Unaligned end が別の領域に十分な量マップすることができれば、そこまでの距離のInsertionやDeletion、構造変異と考えられます。
- 注意：このツールでは、同一染色体内の構造変異のみが検出可能です。



- 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。

Ecoli.FLX (single) (BP)
Left breakpoint/Right breakpoint annotations (334)

Ecoli.FLX (single) (InDel) Variants

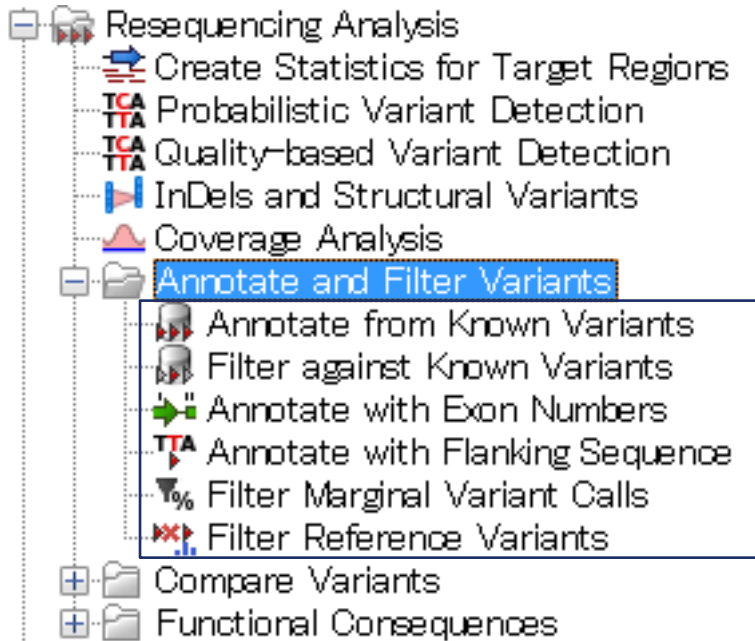
Ecoli.FLX (single) (SV)
Deletion/Insertion/Inversion/Replacement annotations (24)

Track List Settings:
 Navigation: NC_010473 (4,686,1...)
 Range: 4034284
 Insertions: Ecoli.l...
 Find:
 Track layout:
 - DNA sequence track
 - CDS track
 - Gene track
 - Reads track

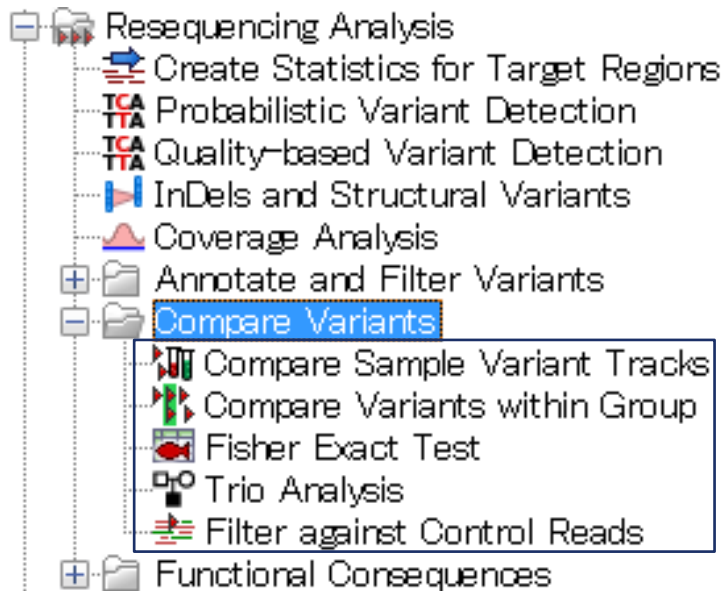
Chromosome	Region	Name	Evidence	Length	Reference sequence
NC_010473	1738503..1740635	Deletion	Cross mapped breakpoints	1351	TGGATTGGUUUATATTTCCAGACATCTGTTATCACTTAACCCATTACAAGCCCGCTGCCGCAGATATCCCGTGGCGAGCC
NC_010473	1793332..1793335	Insertion	Close breakpoints	0	
NC_010473	1961625..1961635	Insertion	Close breakpoints	0	
NC_010473	2327444..2327445	Replacement	Paired breakpoint	21	GA
NC_010473	2471526..2472293	Deletion	Cross mapped breakpoints	768	GGTGATGCTGCCAACTTACTGATTTAGTGTATGATGGTGTTTTTGAGGTGCTCCAGTGGCTTCTGTTTCTATCAGCTGTCCCT
NC_010473	2472293..2472294	Insertion	Close breakpoints	0	
NC_010473	2895595..2895598	Insertion	Close breakpoints	0	
NC_010473	3003961..3005291	Deletion	Cross mapped breakpoints	1331	TGGATTGGCCCTATATTTCCAGACATCTGTTATCACTTAACCCATTACAAGCCCGCTGCCGCAGATATCCCGTGGCGAGCC
NC_010473	3170620..3171813	Deletion	Cross mapped breakpoints	1194	GAAGGTGCGAACAAAGTCCCTGATATGAGATCATGTTTGTCACTCGGAGCCATAGAACAGGGTTCATCATGAGTCATCAACTT
NC_010473	complement(3199470..3211928)	Inversion	Cross mapped breakpoints	12459	TGATGAATCCCTAATGATTTTGGTAAAATCATTAAAGTTAAGGTGGATACACATCTTGTCATATGATCAAATGGTTTCGCGA
NC_010473	complement(3200798..3213256)	Inversion	Cross mapped breakpoints	12459	GCGCGTAAAGTTTTCCGGCTCAACAAGAGAAGGCGCTTCCGTAATGTAAGCCAGCTCGTTTTATCCAGGCGCGTTTTCCG
NC_010473	4034585..4035825	Deletion	Cross mapped breakpoints	1331	TGGATTGGCCCTATATTTCCAGACATCTGTTATCACTTAACCCATTACAAGCCCGCTGCCGCAGATATCCCGTGGCGAGCC
NC_010473	4166331..4167088	Deletion	Cross mapped breakpoints	768	GGTGATGCTGCCAACTTACTGATTTAGTGTATGATGGTGTTTTTGAGGTGCTCCAGTGGCTTCTGTTTCTATCAGCTGTCCCT
NC_010473	4387176..4387943	Deletion	Cross mapped breakpoints	768	GGTGATGCTGCCAACTTACTGATTTAGTGTATGATGGTGTTTTTGAGGTGCTCCAGTGGCTTCTGTTTCTATCAGCTGTCCCT
NC_010473	4551091..4552531	Deletion	Cross mapped breakpoints	1441	TACTGCACCCATTTTGTGGACGATGAAATGGAATAGCCCTAATATGTCAAAGCCAAAATACCCTTTTGAAAAGCGCCTTG/

Filter

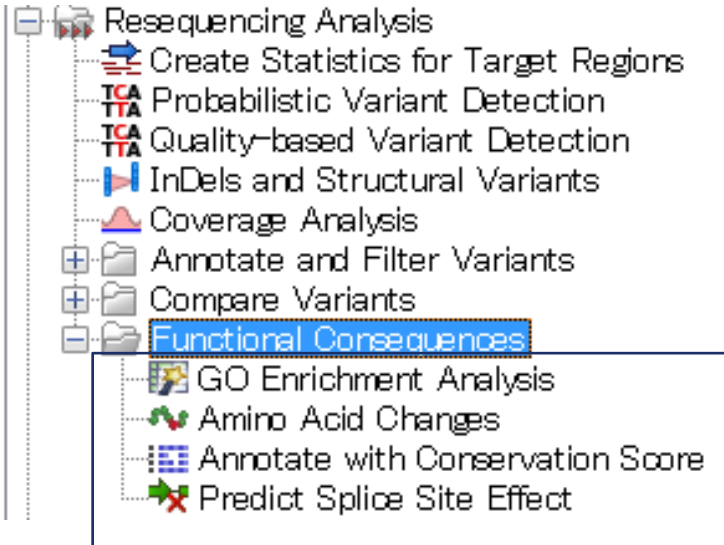
Create Track from Selection



- 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

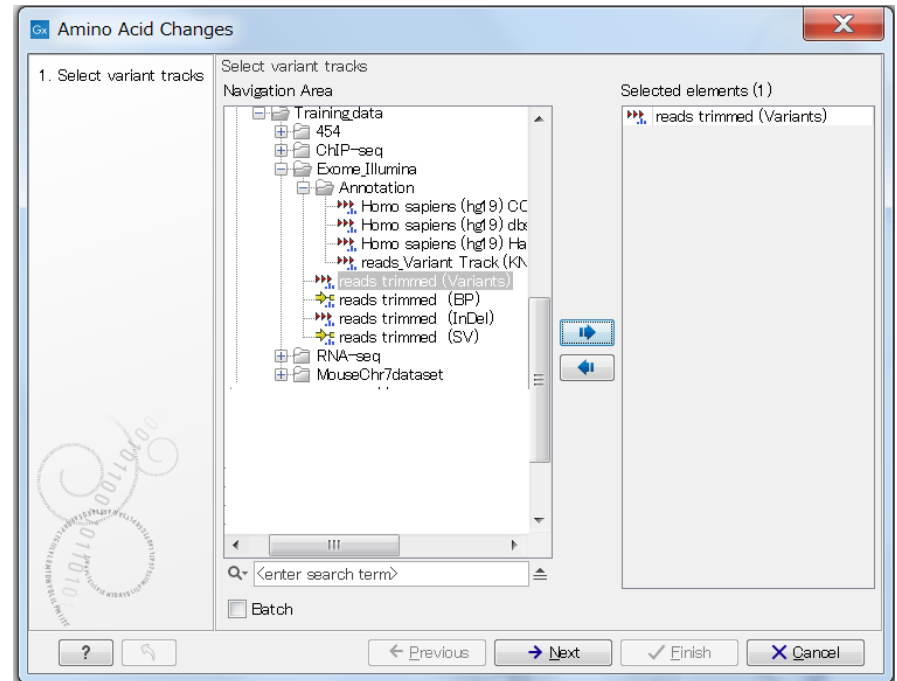
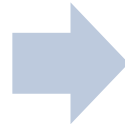
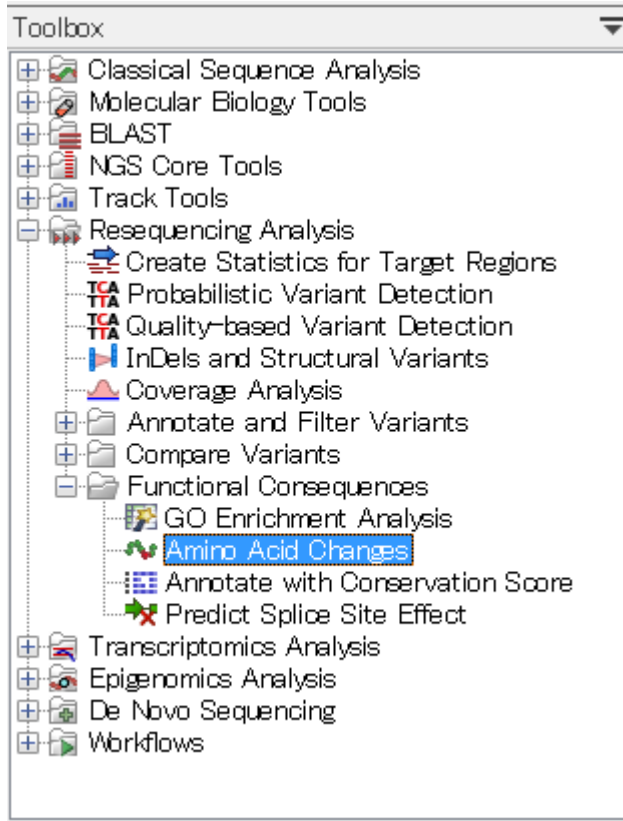


- Compare Sample Variant Tracks: 2つの variant track を比較して、共通する、または、異なる variant を出力する (Ver6.5で追加)。
- Compare Variants within Group : グループの中で common variants を検索する。Frequency を % で指定できる。
- Fisher Exact Test : Case-control study で、case に有意に存在する variants を検出する。
- Trio Analysis : 子供と両親のデータを用いて trio 解析を行う。Variants 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。

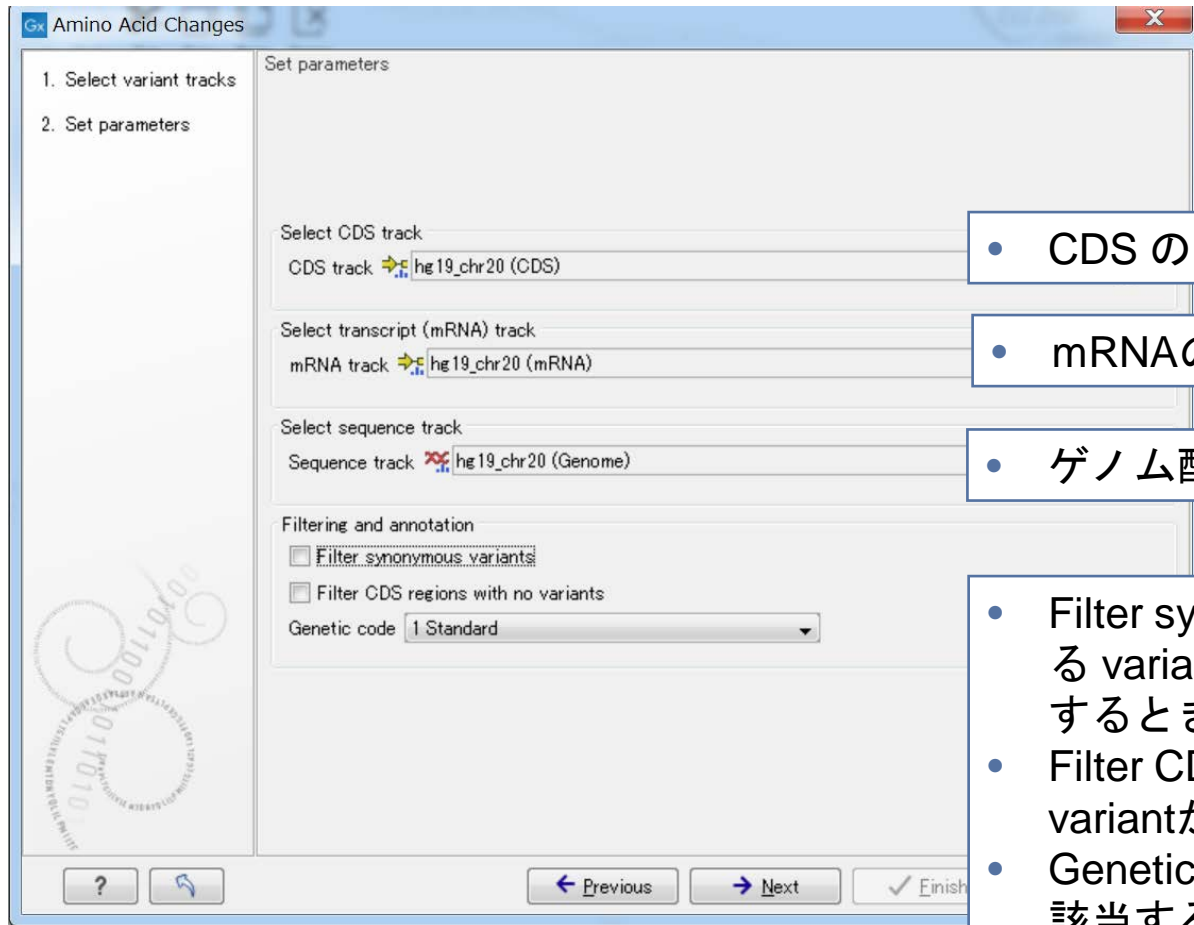


- GO Enrichment Analysis : 検出された variants が含まれる遺伝子にどのような Gene Ontology と関連するものが多いのかを解析する。
- Amino Acid Changes : variants に、アミノ酸置換に関するアノテーション付けを行う。
- Annotate with Conservation Score : 異なる種におけるアミノ酸の保存の程度に関する情報をアノテーション付けする。保存の度合いが高いほど、機能的に重要であると期待される。
- Predict Splice Site Effect : variants の splice site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

アミノ酸置換の解析



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



- CDS の track を選択

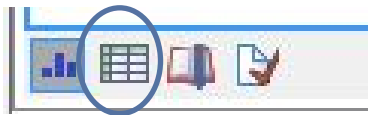
- mRNAのtrackを選択

- 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

Variant Track

reads trimmed (Variants, AAC)



Rows: 517 Table view: Genome Filter:

Linkage	Zygoty	Count	Coverage	Frequency	Forward re...	Reverse re...	Forward/re...	Average qu...	Coding regi...	Amino acid change	Non-syn...
	Homozygous	68	68	100.00	34	34	0.50	29.38	ENST00000...	ENSP00000367436p.[Leu140Val]	ENSP0... Yes
	Homozygous	12	13	92.31	8	4	0.33	23.67	ENST00000...	ENSP00000202816p.[Ile550Thr]	Yes
	Heterozygous	19	47	40.43	18	1	0.05	30.42	ENST00000...	ENSP00000217246p.[Thr58Ile]	ENSP000... Yes
	Heterozygous	21	43	48.84	2	19	0.10	23.71	ENST00000...	ENSP00000367242p.[Asn345Ser]	ENSP... Yes
	Heterozygous	14	24	58.33	4	10	0.29	27.50	ENST00000...	ENSP00000367242p.[Met253Thr]	ENSP0... Yes
	Homozygous	10	10	100.00	4	6	0.40	28.00	ENST00000...	ENSP00000401206p.[Leu384His]	ENSP0... Yes
	Heterozygous	5	10	50.00	3	2	0.40	28.60	ENST00000...	ENSP00000246043p.[Gln446Pro]	ENSP0... Yes
	Homozygous	40	40	100.00	18	22	0.45	28.88	ENST00000...	ENSP00000246043p.[Ser382Ala]	ENSP0... Yes
	Homozygous	29	30	96.67	23	6	0.21	24.55	ENST00000...	ENSP00000392318p.[Val400Gly]	ENSP0... Yes
	Homozygous	52	53	98.11	20	32	0.38	27.29	ENST00000...	ENSP00000400897p.[Ser196Thr]	ENSP0... Yes
	Homozygous	23	24	95.83	12	11	0.48	24.26	ENST00000...	ENSP00000366645p.Pro99Ser	Yes
	Heterozygous	8	16	50.00	4	4	0.50	26.62	ENST00000...	ENSP00000311027p.[Ala6Thr]	ENSP000... Yes
	Homozygous	20	20	100.00	10	10	0.50	24.40	ENST00000...	ENSP00000345553p.[Val323Ile]	ENSP0... Yes
	Heterozygous	39	111	35.14	19	20	0.49	30.26	ENST00000...	ENSP00000255008p.Phe321Ser	Yes
	Homozygous	58	57	98.95	28	29	0.50	32.60	ENST00000...	ENSP00000233006p.[Gln325P...	ENSP0... Yes

Rows: 1,085 Table view: Genome Filter

Reverse rea...	Forward/rev...	Average quality	Coding region change	Amino acid change	Amino acid change in longest transcript	Coding region change
59	0.48	20.03				
46	0.48	19.60	ENST00000341420:c.[1257C>T]; ENST00000217246:c.[272...			ENST00000341420:c.1
32	0.48	21.05				
24	0.47	18.02	ENST00000341420:c.[1200C>A]; ENST00000310348:c.[272...	ENSP00000339912p.[His400Gln]; ENSP000000...	ENSP00000339912p.His400Gln	ENST00000341420:c.1
0	0.00	24.71				
0	0.00	22.43	ENST00000402914:c.[29-80T>C]; ENST00000310348:c.[72...			ENST00000310348:c.7
1	0.05	24.53				
0	0.00	24.12	ENST00000217246:c.[728-76A>G]; ENST00000310348:c.[7...			ENST00000310348:c.7
0	0.00	20.91				
0	0.00	22.88	ENST00000378058:c.[281-45A>G]; ENST00000402914:c.[2...			ENST00000310348:c.5
0	0.00	23.74				
0	0.00	23.75	ENST00000402914:c.[356-76A>G]; ENST00000310348:c.[1...			ENST00000310348:c.1
0	0.00	21.89				
0	0.00	26.40				
0	0.00	21.11				

Show column

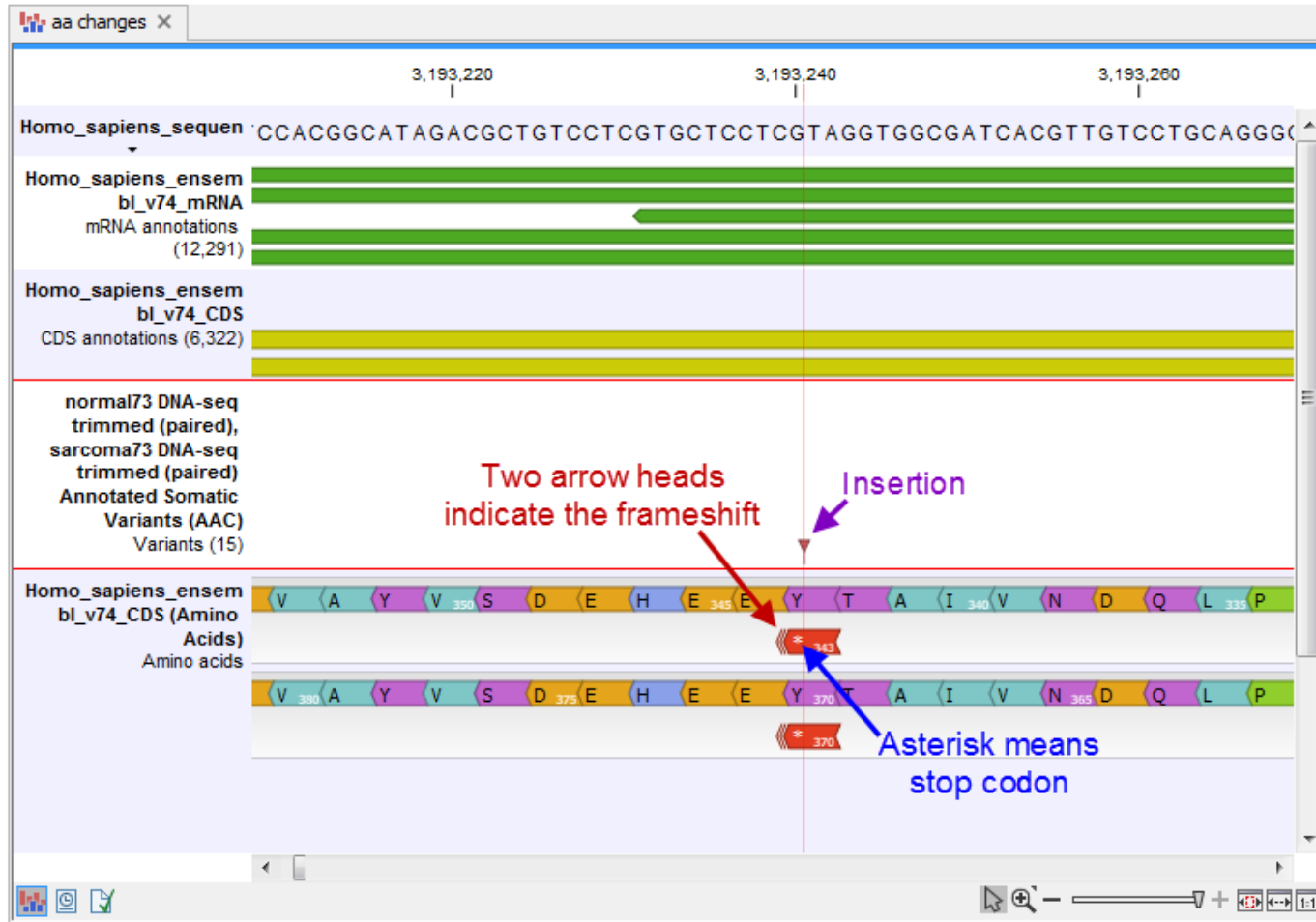
- Chromosome
- Region
- Type
- Reference
- Allele
- Reference allele
- Length
- Linkage
- Zygosity
- Count
- Coverage
- Frequency
- Probability
- Forward read count
- Reverse read count
- Forward/reverse balance
- Average quality
- Coding region change
- Amino acid change
- Amino acid change in longest transcript
- Coding region change in longest trans...
- Other variants within codon
- Non-synonymous

Select All

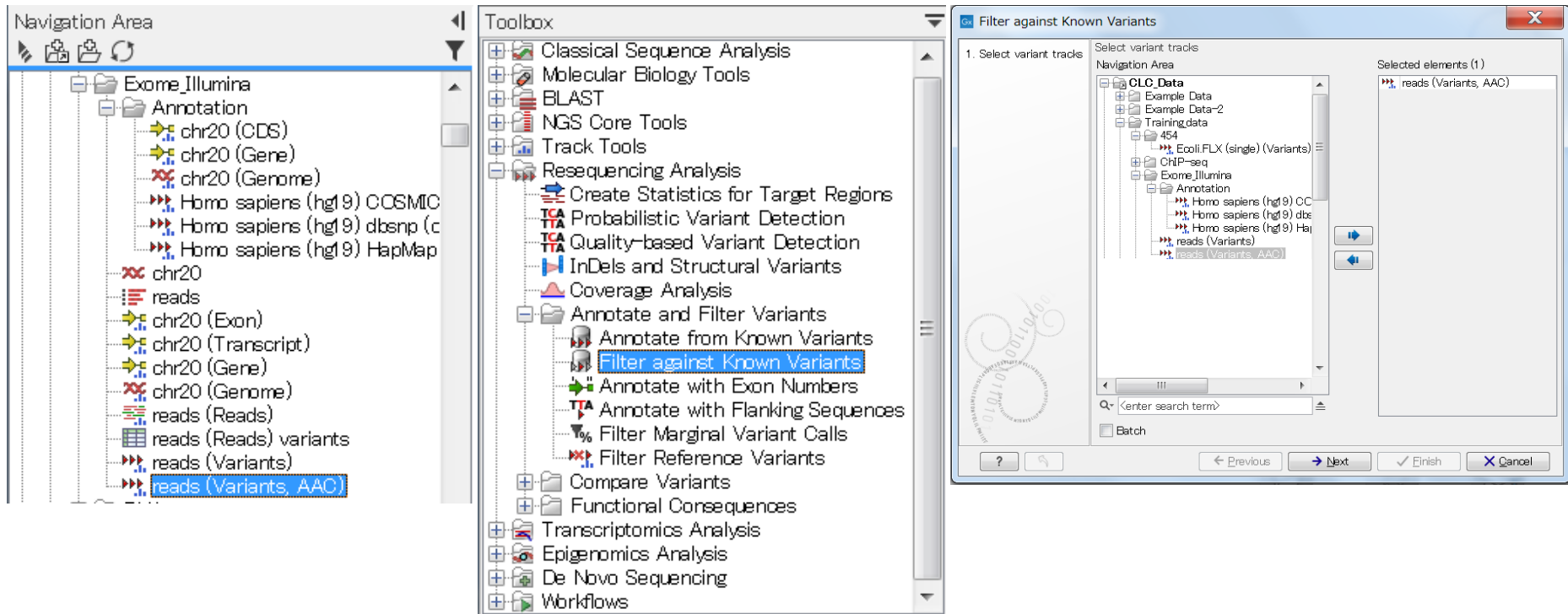
Deselect All

- 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

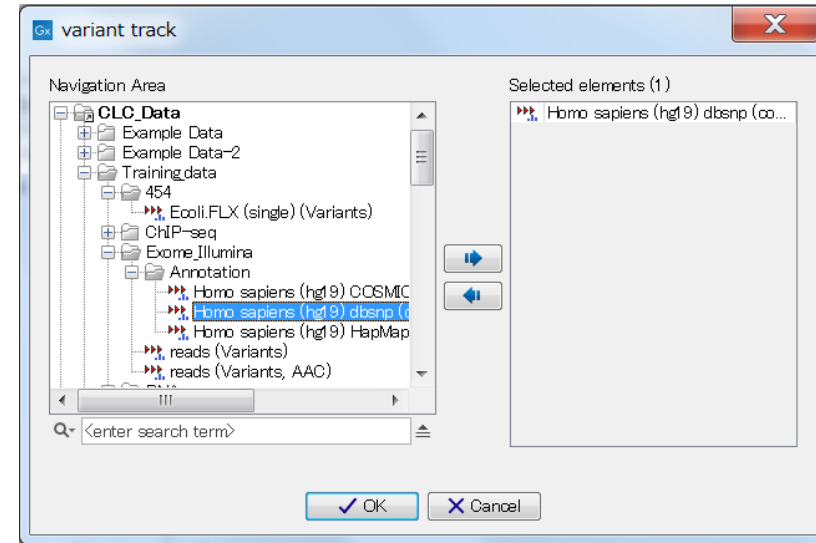
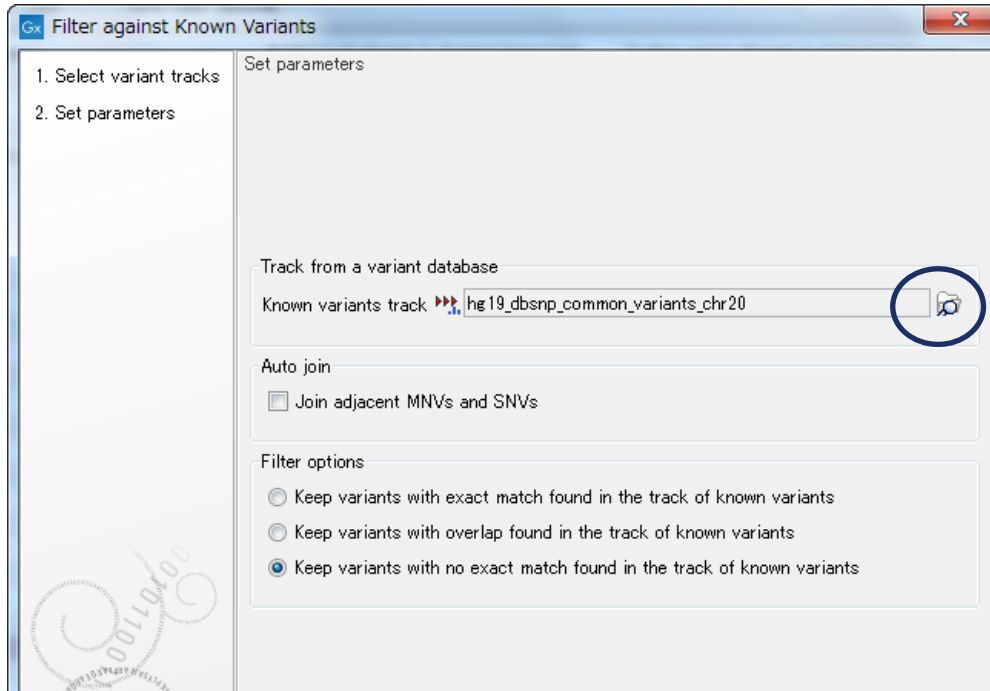


変異のフィルタリング



The screenshot displays the software's interface for selecting a workflow. On the left, the 'Navigation Area' shows a tree structure of data tracks, with 'reads (Variants, AAC)' selected. In the center, the 'Toolbox' lists various analysis tools, with 'Filter against Known Variants' highlighted under the 'Annotate and Filter Variants' category. On the right, the 'Filter against Known Variants' wizard dialog box is open, showing the same 'Navigation Area' tree with 'reads (Variants, AAC)' selected in the 'Selected elements (1)' list.

- 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。



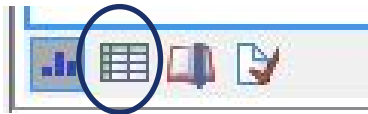
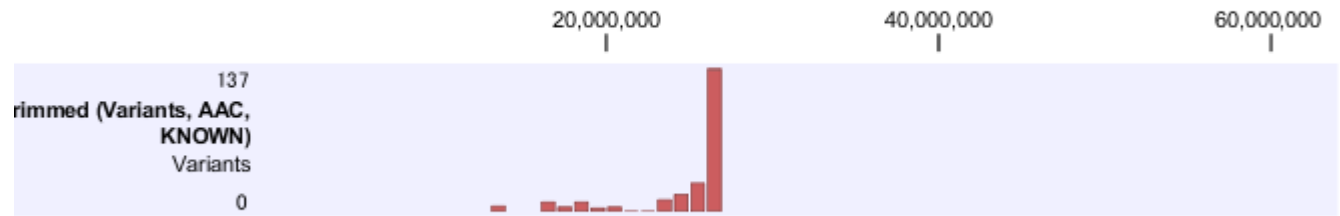
- Known variants track: 比較したい変異トラックを選択。
- Auto join: 隣り合わせの変異について、フィルターをかける際に一つの変異として扱うかどうか。
- Filter Option
 - Matchを残す : アレルまで完全に一致しているものを残す。
 - Overlapを残す : オーバーラップがあるものを残す。
 - Not matchを残す : 完全に一致しなかったものを残す。

- SNPのトラックを選択。

reads trimmed (Variants, AAC, KNOWN)

To show more tracks together, create a track list.

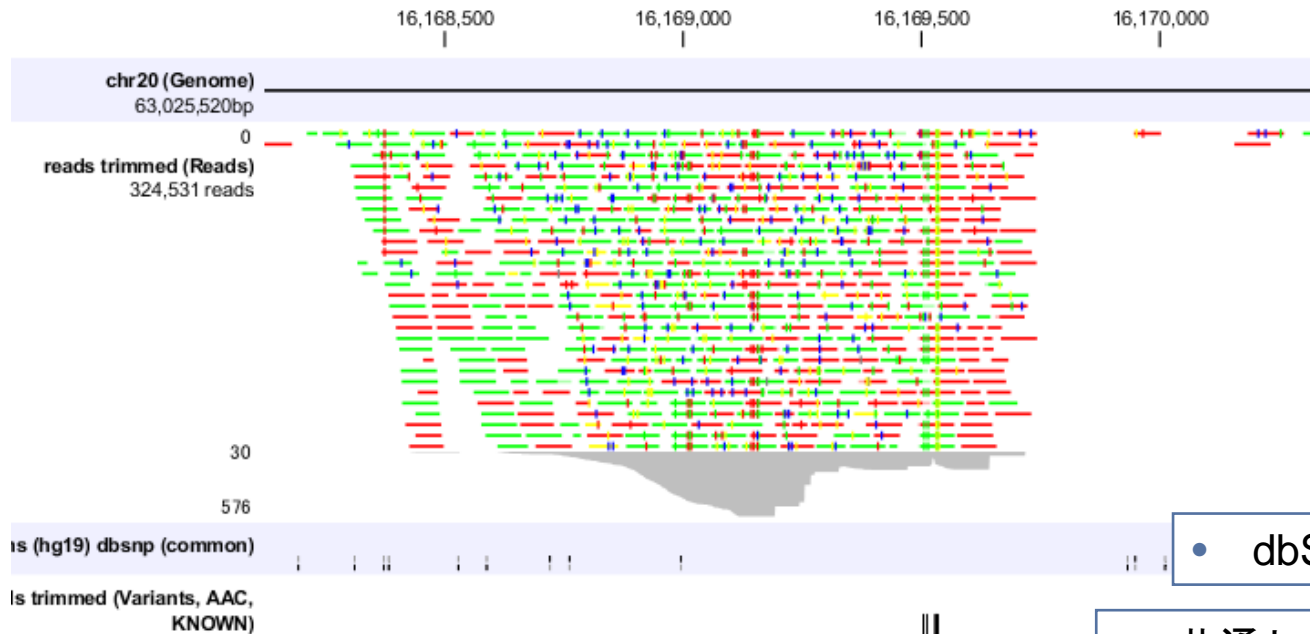
Create Track List



Rows: 236 Table view: Genome Filter:

Chromoso...	Region	Type	Reference	Allele	Reference ...	Linkage	Zygoty	Count	Coverage	Fre
chr20	13829044	SNV	A	C	No		Homozygous	70	70	
chr20	13829044	SNV	C	T	No		Homozygous	55	55	
chr20	13829136	SNV	A	G	No		Heterozygous	5	10	
chr20	13845987	SNV	C	T	No		Homozygous	55	55	
chr20	13847420	SNV	A	G	No		Homozygous	80	81	
chr20	16169504	SNV	C	T	No		Homozygous	24	25	
chr20	16169512	SNV	A	T	No		Homozygous	30	30	
chr20	16169529	SNV	A	G	No		Homozygous	76	76	
chr20	16169532	SNV	G	T	No		Homozygous	73	73	
chr20	16169535	SNV	A	G	No		Homozygous	70	70	

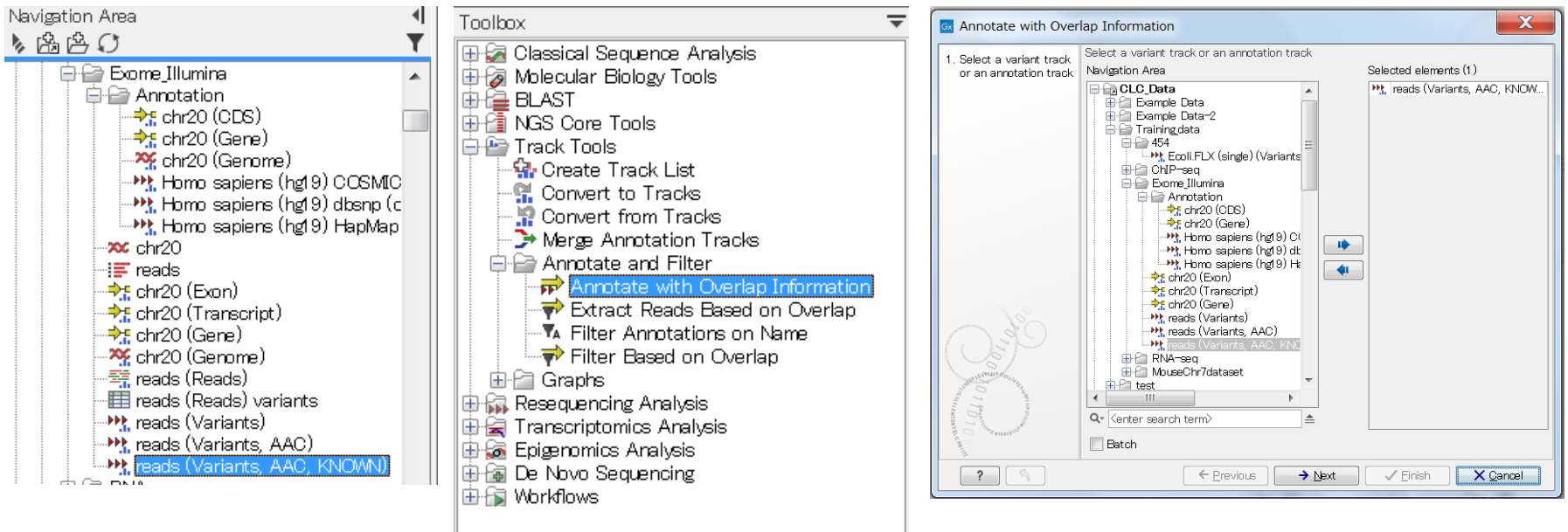
トラックで表示



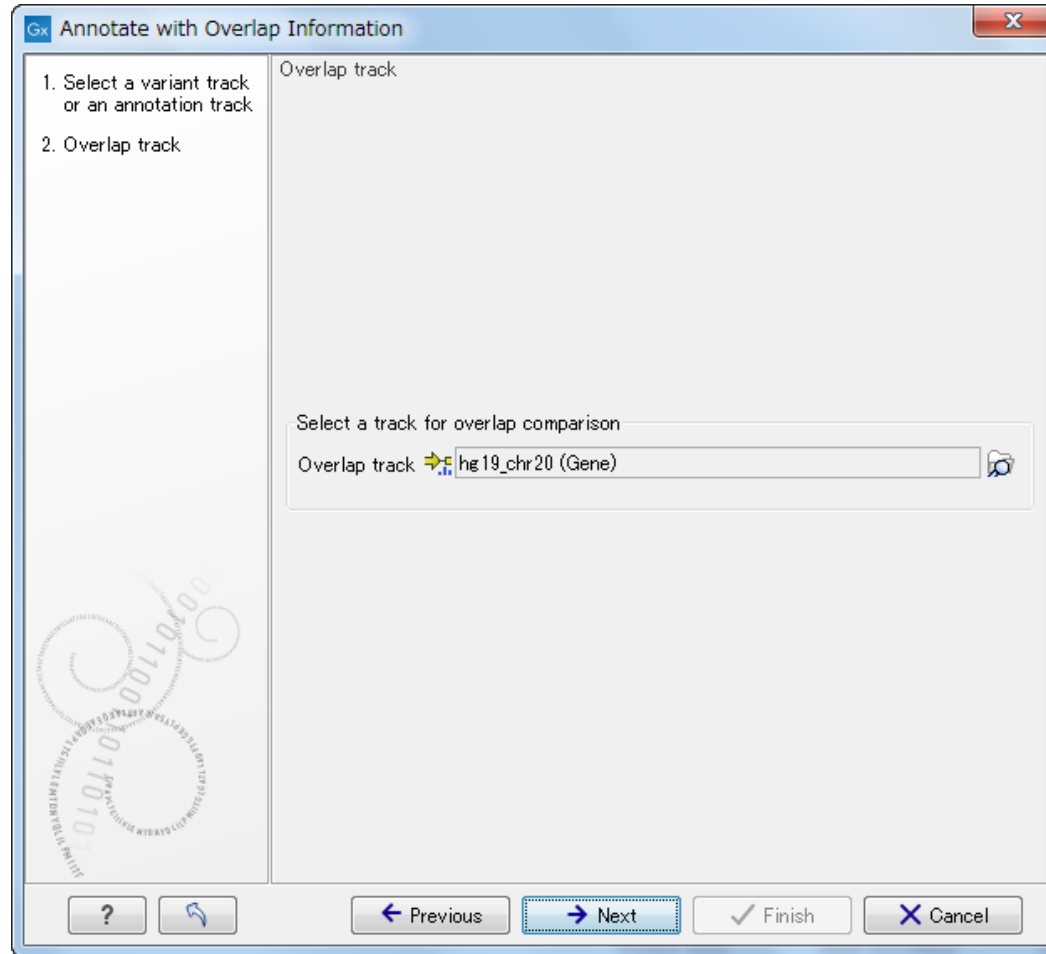
• dbSNP

• 共通しなかったもの

遺伝子のアノテーション

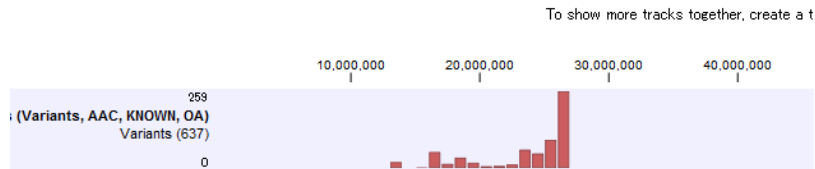


- Navigation Areaから変異トラック(フィルタリングしたもの)を選択。
- Toolboxから Track tools > Annotate and Filter > Annotate with Overlap Information を選択、ダブルクリック。
- ウィザードが起動し、選択したデータが選ばれていることを確認。



- Gene の track を選択。

reads trimmed (Variants, AAC, KNOWN, OA)



transcript	Coding region change in longest transcript	Other varian...	Non-synony...	hg19_chr20 (Gene)	Gene Cards	ENSEMBL
ENST00000399002	c.289A>C	No	No	SPTLC3	SPTLC3	ENSG00000172296
ENST00000399002	c.303+84C>T	No	-	SPTLC3	SPTLC3	ENSG00000172296
ENST00000399002	c.303+84C>T	No	No	SPTLC3	SPTLC3	ENSG00000172296
ENST00000399002	c.1467A>G	No	Yes	SPTLC3	SPTLC3	ENSG00000172296
ENST00000262487	c.80C>G	No	-	-	-	-
ENST00000262487	c.80C>G	No	Yes	ISM1	ISM1	ENSG00000101230
ENST00000262487	c.80C>G	No	No	ISM1	ISM1	ENSG00000101230
ENST00000337743	c.1105C>T	No	No	TASP1	TASP1	ENSG00000089123
ENST00000337743	c.1105C>T	No	No	TASP1	TASP1	ENSG00000089123
ENST00000337743	c.1100A>G	No	Yes	TASP1	TASP1	ENSG00000089123
ENST00000284951	c.1570+8G>A	No	-	-	-	-
ENST00000284951	c.1570+8G>A	No	-	SEL1L2	SEL1L2	ENSG00000101251
ENST00000284951	c.1332T>C	No	No	SEL1L2	SEL1L2	ENSG00000101251
ENST00000310348	c.728+80T>C	No	-	MACROD2	MACROD2	ENSG00000172264
ENST00000310348	c.728+76A>G	No	-	MACROD2	MACROD2	ENSG00000172264



www.genecards.org/cgi-bin/carddisp.pl?gene=SPTLC3#pathways_interactions

SPTLC3 Gene
protein-coding **GIFS: 53**
SCID: GC2NP012938

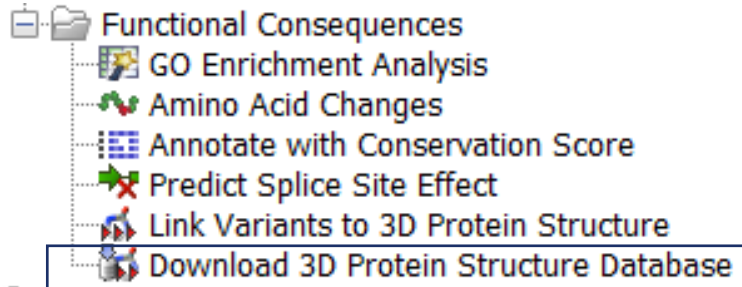
Serine Palmitoyltransferase, Long Chain Base Subunit 3
(Previous names: chromosome 20 open reading frame 38, serine palmitoyltransferase,...)
(Previous symbols: C20orf38, SPTLC2L)

Aliases for SPTLC3 gene
(According to ¹HGNC, ²Entrez Gene, ³UniProtKB/Swiss-Prot, ⁴UniProtKB/TrEMBL, ⁵OMIM, ⁶GeneLoc, ⁷Ensembl, ⁸OMIM, ⁹miRBase, ¹⁰RNAdb, ¹¹HMDB, ¹²NCBI, ¹³NONCODE and/or ¹⁴RNAdb)

Aliases
Serine Palmitoyltransferase, Long Chain Base Subunit 3^{1,2}
SPTLC2L^{1,2,3,5}
C20orf38^{2,3}
Serine Palmitoyltransferase, Long Chain Base Subunit 2-Like
(Aminotransferase 2)^{1,2}
Long Chain Base Biosynthesis Protein 26^{1,2}
Long Chain Base Biosynthesis Protein 3^{1,2}
Serine-Palmitoyl-CoA Transferase 3^{1,2}
LCB 3^{1,2}
SPT 3^{1,2}
EC 2.3.1.50^{1,8}

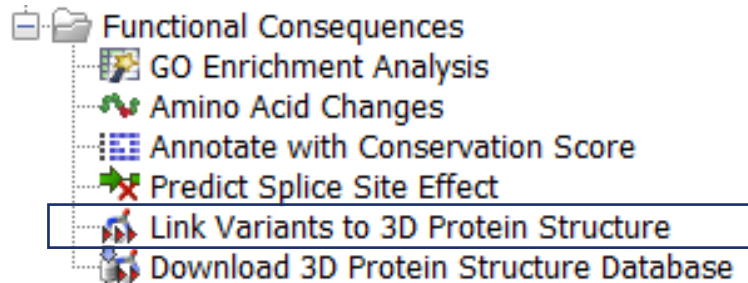
External IDs: HGNC: 16253¹ Entrez Gene: 55304² Ensembl: ENSG00000172296⁶ OMIM: 611120⁵ UniProtKB: Q9JUV7³

3 D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。



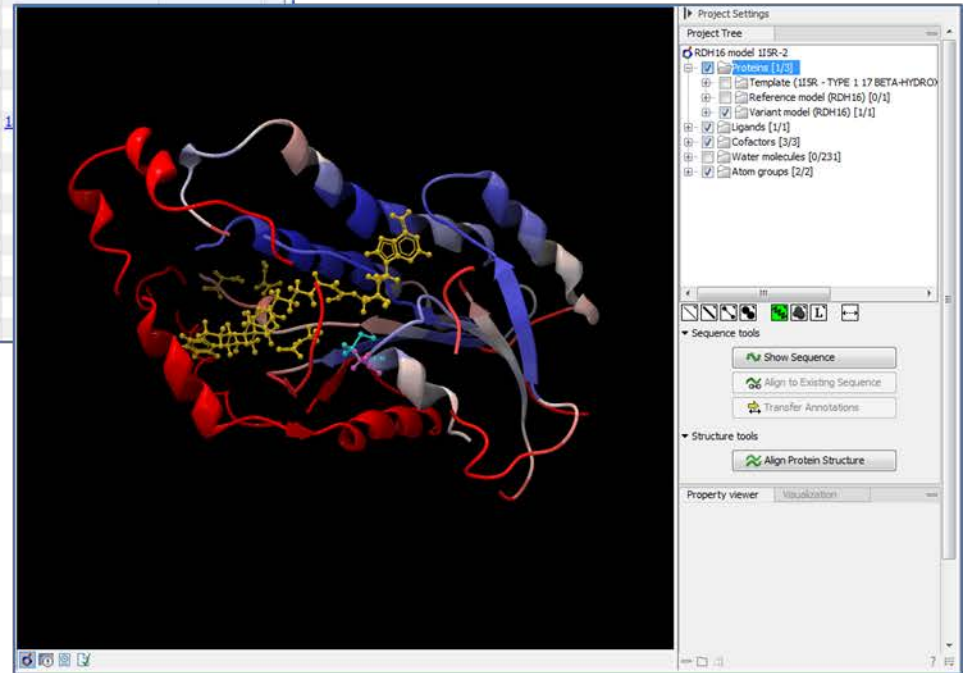
Link Variants to 3D Protein Structure:

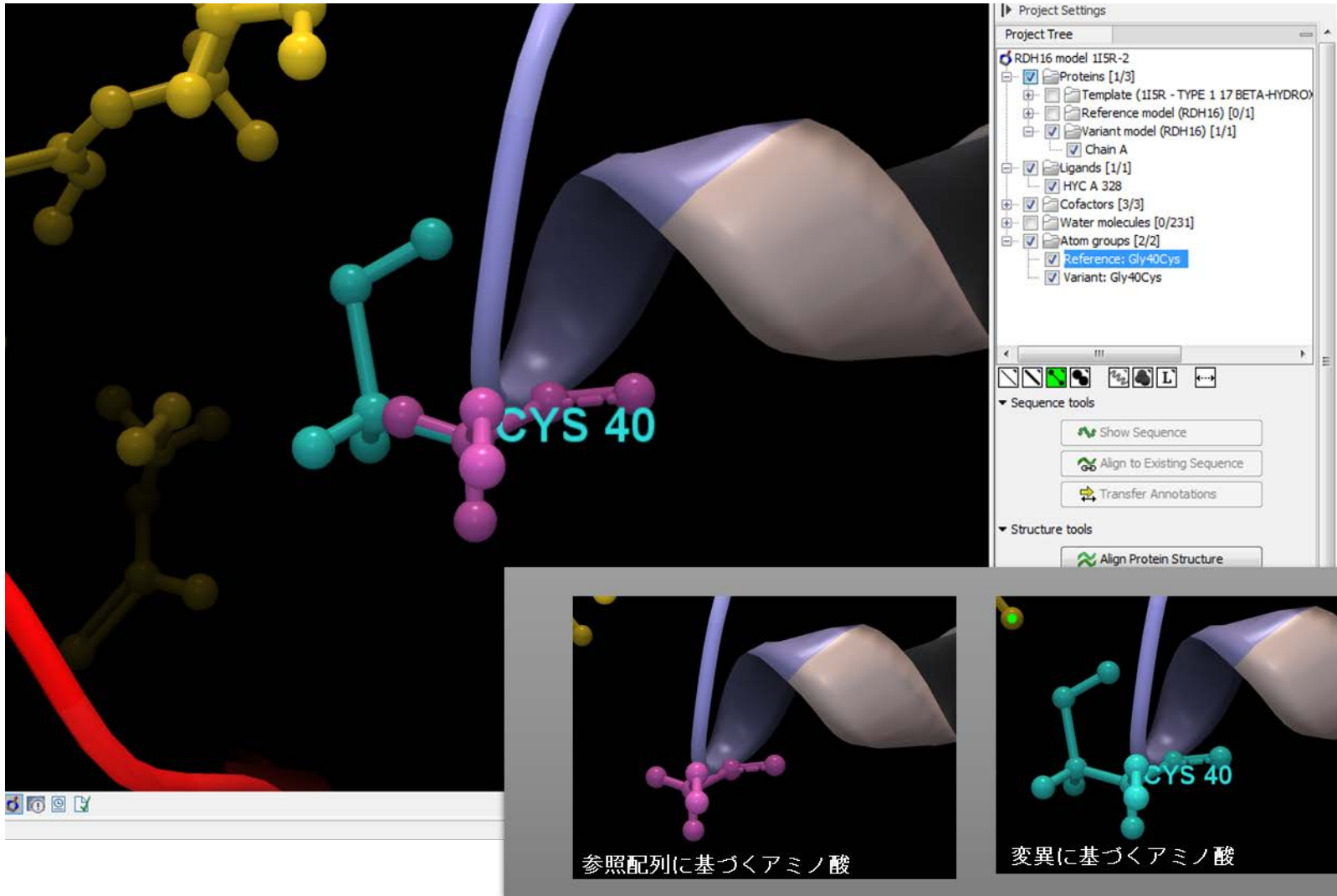
ダウンロードしたアミノ酸配列に対して、変異の検出からアミノ酸置換させた配列を使いBLAST検索をかけます。

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	MME1 Met518Thr			
			0.78	ACTRT2 (synonymous)			
			0.97	ARHGEF16 Ser360Arg			
			0.00	LRRC47 (synonymous)			
			0.95	CEP104 (nonsense)			





Project Settings

Project Tree

- RDH16 model 115R-2
 - Proteins [1/3]
 - Template (115R - TYPE 1 17 BETA-HYDRO)
 - Reference model (RDH16) [0/1]
 - Variant model (RDH16) [1/1]
 - Chain A
 - Ligands [1/1]
 - HYC A 328
 - Cofactors [3/3]
 - Water molecules [0/231]
 - Atom groups [2/2]
 - Reference: Gly40Cys
 - Variant: Gly40Cys

Sequence tools

- Show Sequence
- Align to Existing Sequence
- Transfer Annotations

Structure tools

- Align Protein Structure

参照配列に基づくアミノ酸

変異に基づくアミノ酸

お疲れ様でした。

