



CLC Genomics Workbench ハンズオントレーニング RNA-seq

株式会社CLCバイオジャパン
シニアフィールドバイオインフォマティクスサイエンティスト
宮本真理 Ph.D.
mmiyamoto@clcbio.co.jp



- support@clcbio.co.jp



アジェンダ

- Genomics Workbench 概要
- 今日のデータ
- RNA-seq解析
 - データインポート
 - QC
 - RNA-seq
 - 発現差解析

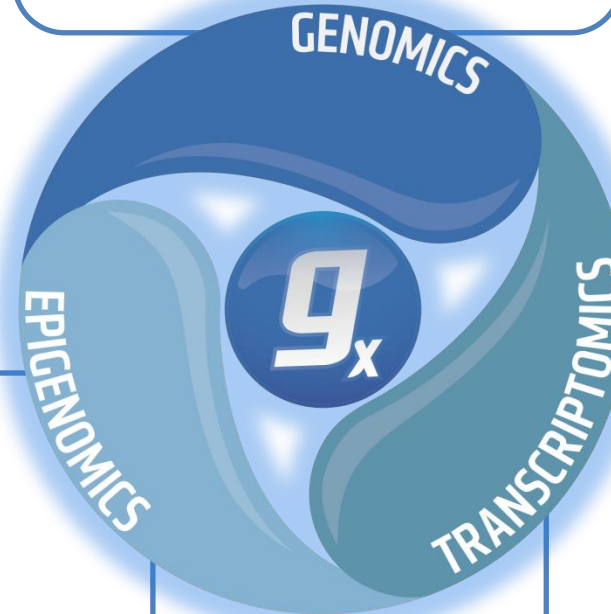
CLC Genomics Workbench 概要



CLC Genomics Workbench



*Genomics:
de novo assembly,
SNP detection*



*Epigenomics:
ChIP-seq,
Peak Finding*

*Transcriptomics:
RNA-seq,
Digital Expression analysis*

解析ワークフロー

新規生物種

変異解析

ChIP-seq

RNA-seq

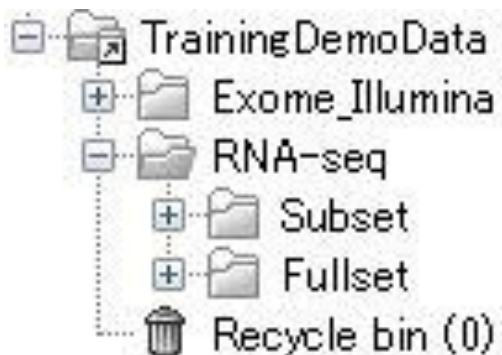
small RNA



今日のデータ

 TrainingDemoData.zip

.zip のままGenomics Workbench へ
インポート

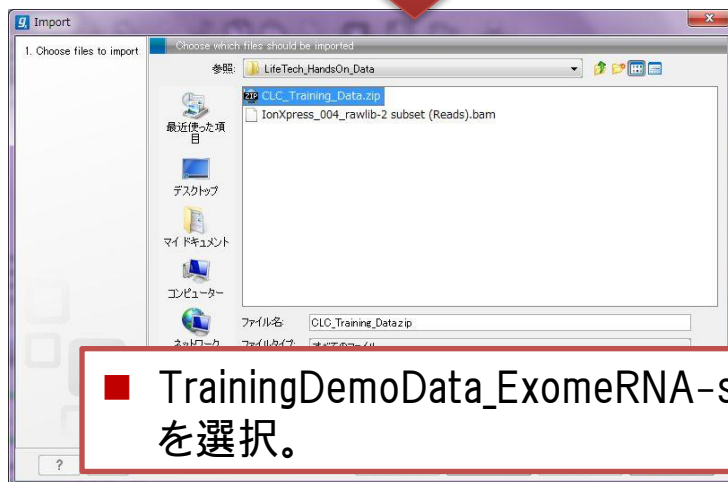
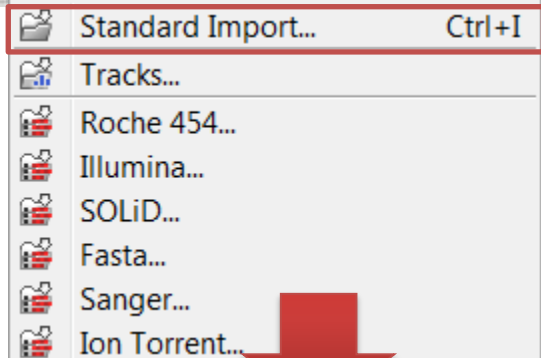


インポートすると変異検出とRNA-seqのデータ

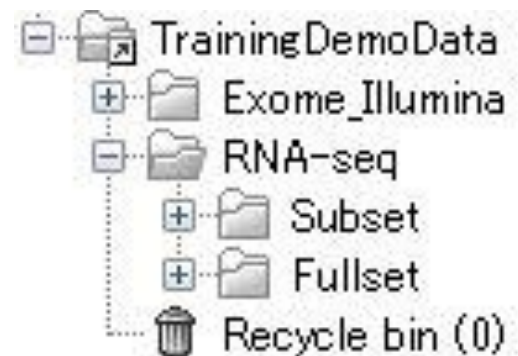


デモデータインポート

- ImportからStandard Import を選択。



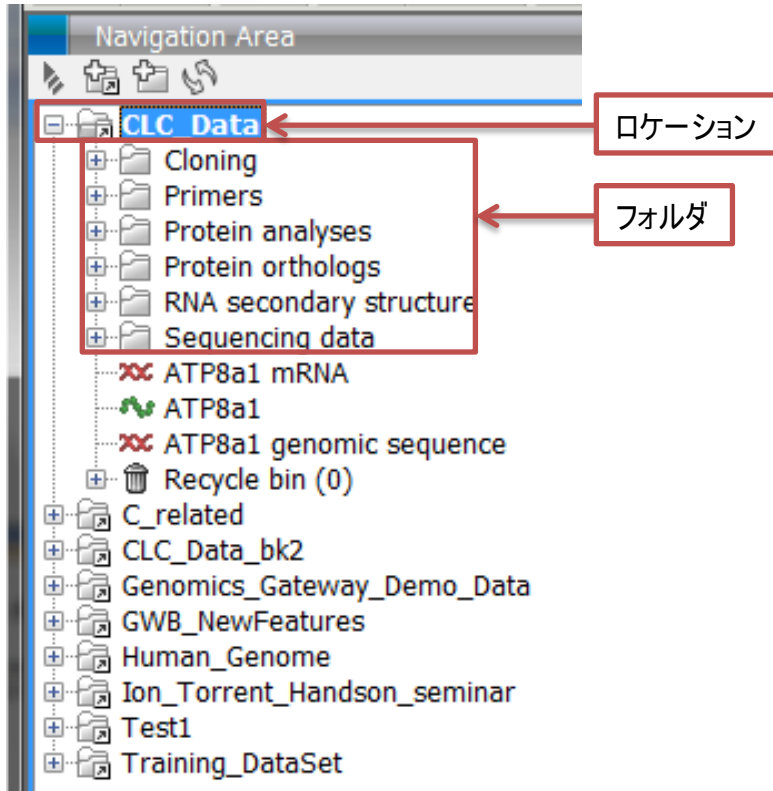
■ TrainingDemoData_ExomeRNA-seq.zip を選択。



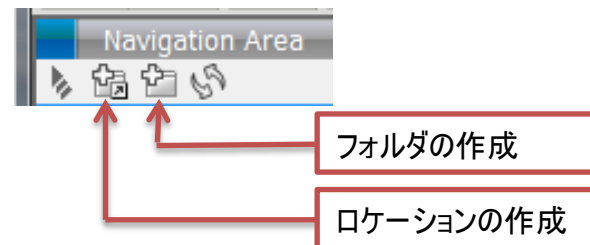
■ TrainingDemoData フォルダが作成

CLC Genomics Workbench 名称と注意事項

LocationとFolder



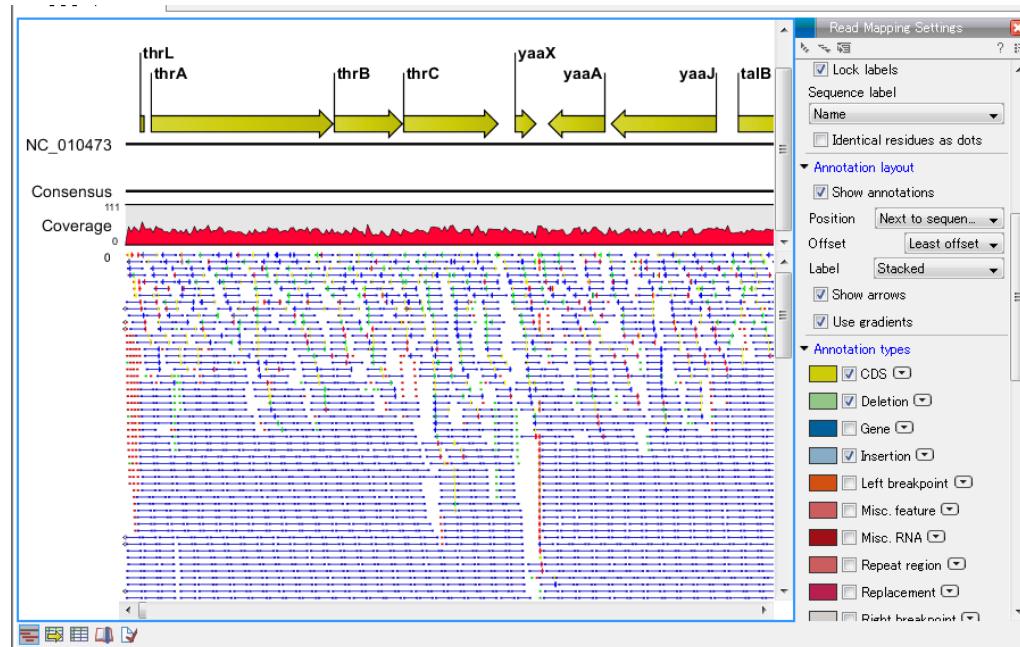
- Genomics Workbenchではデータを以下のような階層構造で保存可能です。フォルダの一番上位の階層を「Location」と呼び、その下の階層を「Folder」と呼びます。
- データの保存場所はロケーション毎に設定可能です。たとえばあるデータはCドライブに保存し、あるデータはDドライブに保存するといった事が可能です。
- ロケーション、フォルダの作成は以下のアイコンから作成できます。



トラックとスタンドアロンフォーマット

- Genomics Workbenchはビューアにスタンドアロンフォーマットとトラックフォーマットがあります。
- 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。

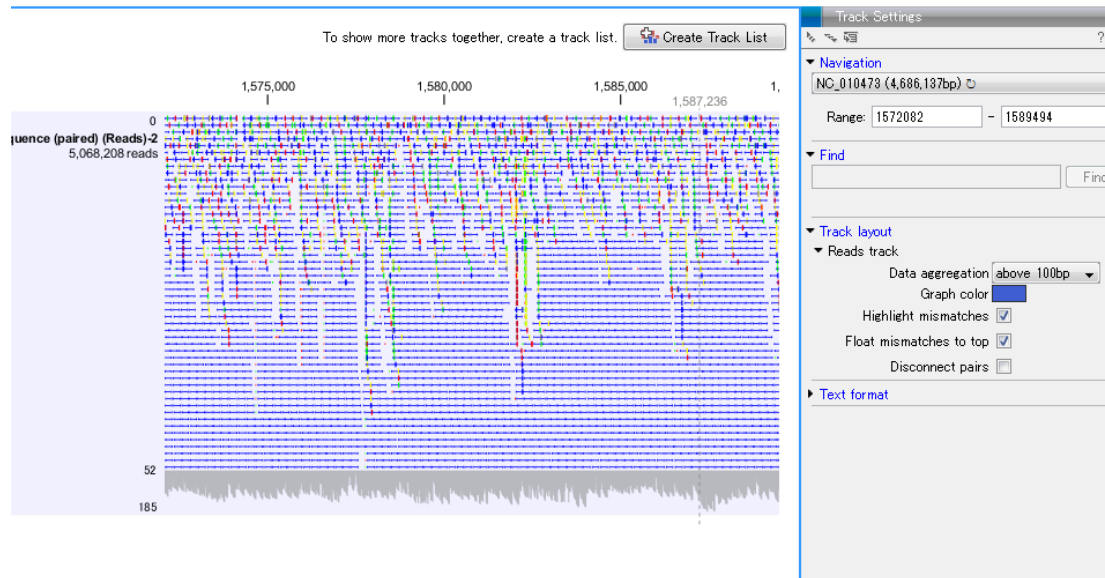
reads mapping



トラックとスタンドアロンフォーマット

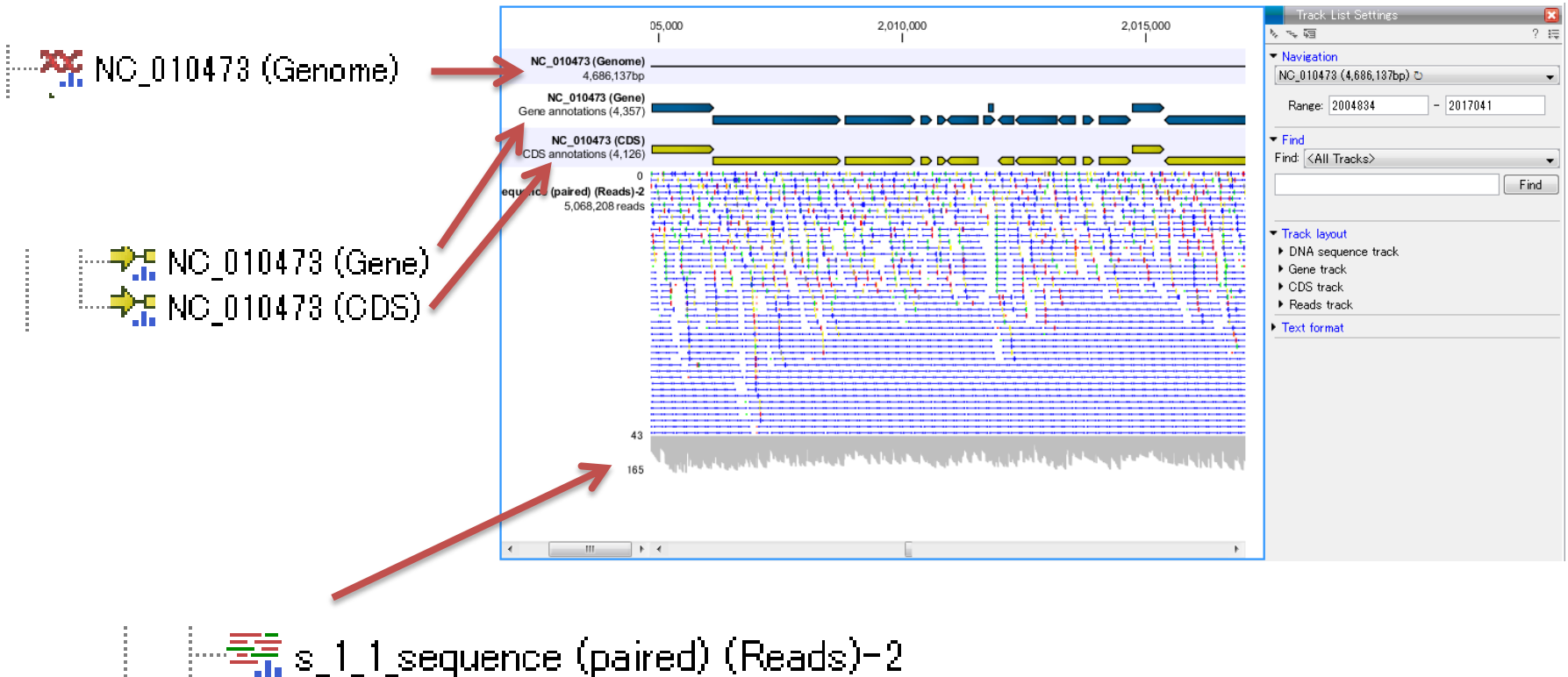
- トラックフォーマットでは、リードやゲノム配列、アノテーションがばらばらのファイルになっており、好きに組み合わせる表示が可能です。

 reads (Reads)





トラックとスタンドアロンフォーマット

- 複数のトラックを組み合わせて好きなビューを作成できます。











トラックと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

-  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 ■ アノテーションTrack
-  Homo sapiens (hg19)_Exon
-  Homo sapiens (hg19)_Gene
-  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。

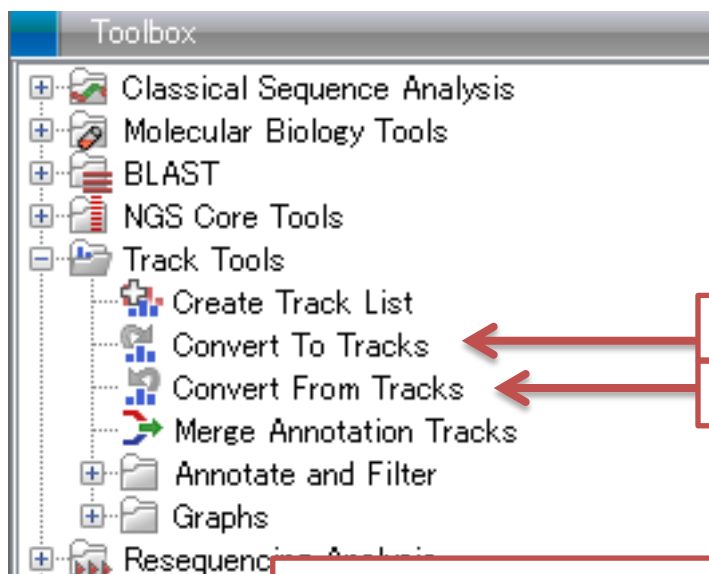
解析によって必要なフォーマット

- 以下に主な解析ツールで必要となるフォーマットをまとめています。解析の際にデータが選べないという場合、必要とするフォーマットに変換されていない場合がありますので、こちらをご参照ください。

解析方法	必要となるフォーマット
マッピング	参照配列はスタンドアロンフォーマット、トラックフォーマットのいずれも可。ただしある領域をマスクしたり、ある領域にのみマッピングさせるような場合、その領域を指定するファイルはトラックフォーマットの必要がある。
ターゲット領域のカバレッジ計算	ターゲット領域はトラックフォーマットの必要がある。
変異検出	変異検出に使うマッピングファイルは、スタンドアロンフォーマット、トラックフォーマットいずれも可。
RNA-seq	参照配列はスタンドアロンフォーマットが必要。またアノテーションとして、GeneとmRNAを含んでいることが必要。トラックから変換する際には、参照配列と、対応するGene、mRNAのアノテーションを選択し、トラックへ変換する必要がある。
ChIP-seq	インプットとなるマッピングファイルはスタンドアロンフォーマットの必要がある。

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Genomics Workbench の Toolbox > Track tools の中のツールを使って変換可能です。



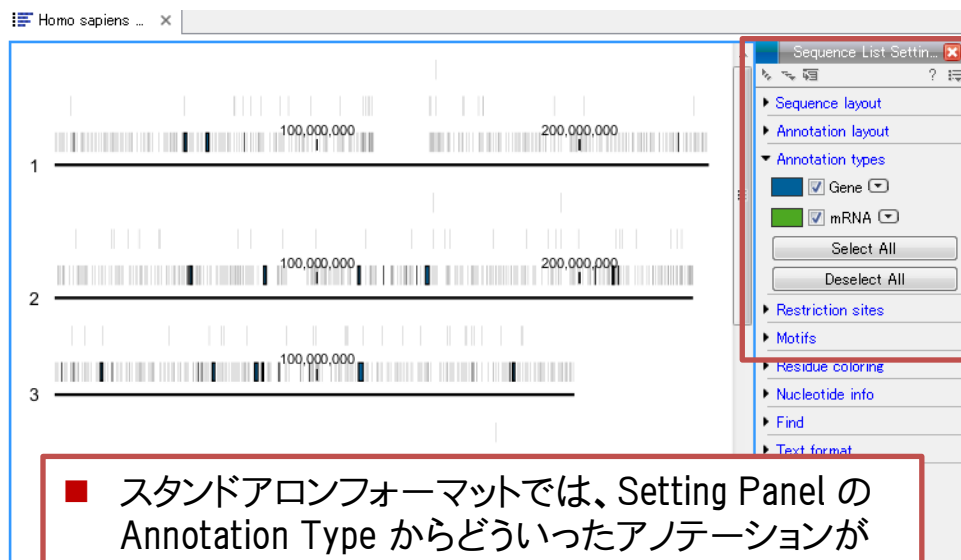
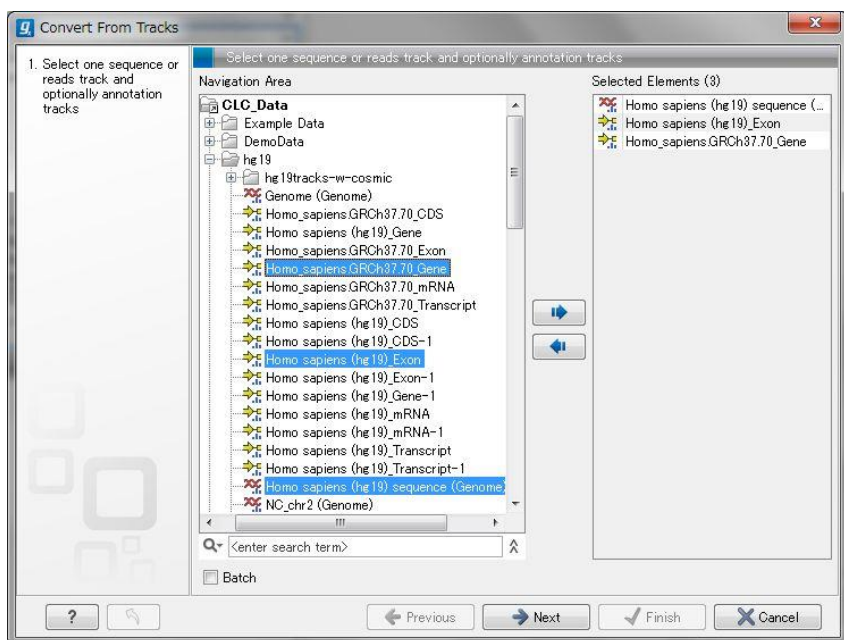
■ スタンドアロンフォーマットからトラックへの変換。

■ 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、スタンドアロン内を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

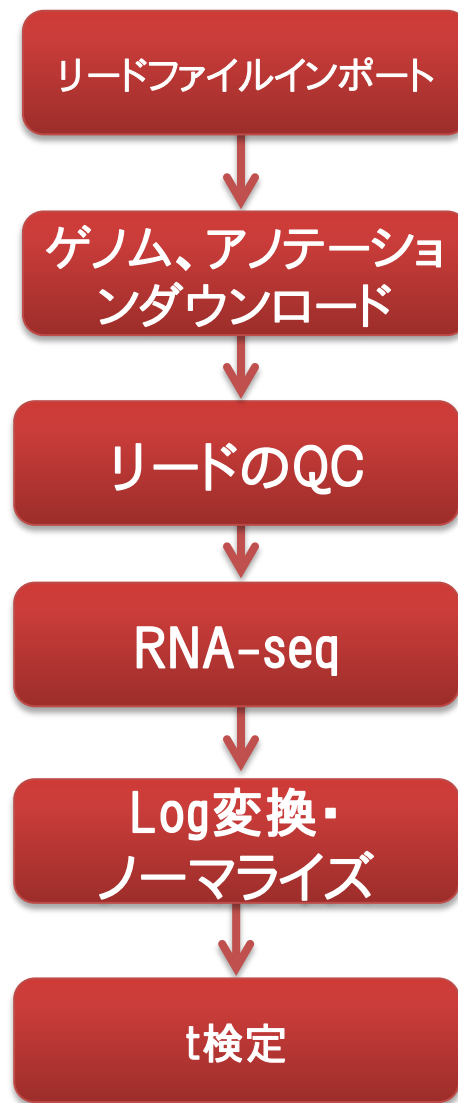
スタンドアロンフォーマットへ変換する場合、スタンドアロン内に含めるアノテーショントラックを含めて変換するようにしてください。



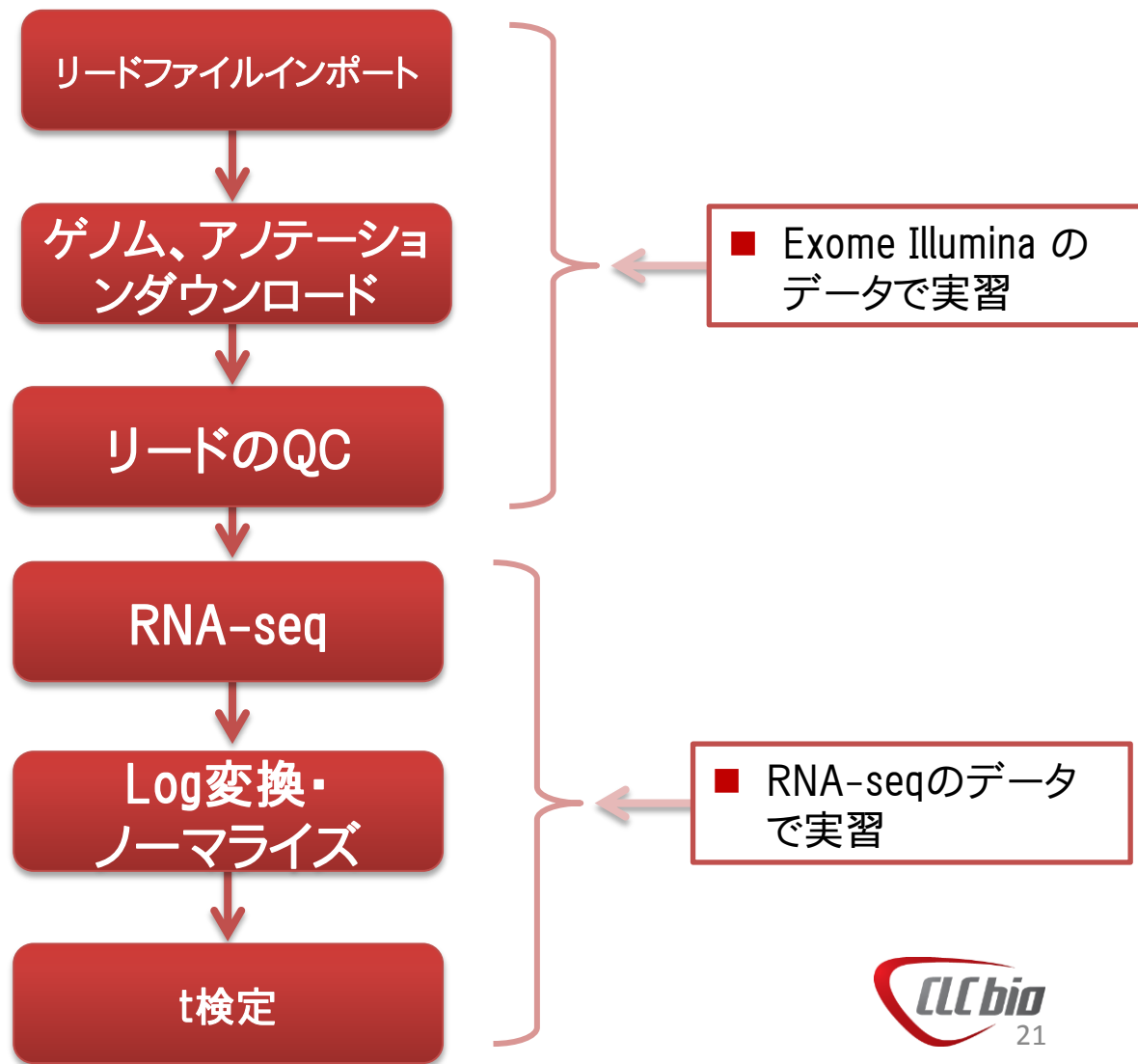
- スタンドアロンフォーマットでは、Setting Panel の Annotation Type からどういったアノテーションが 付属しているか確認できます。

RNA-seq 解析

変異検出の流れ

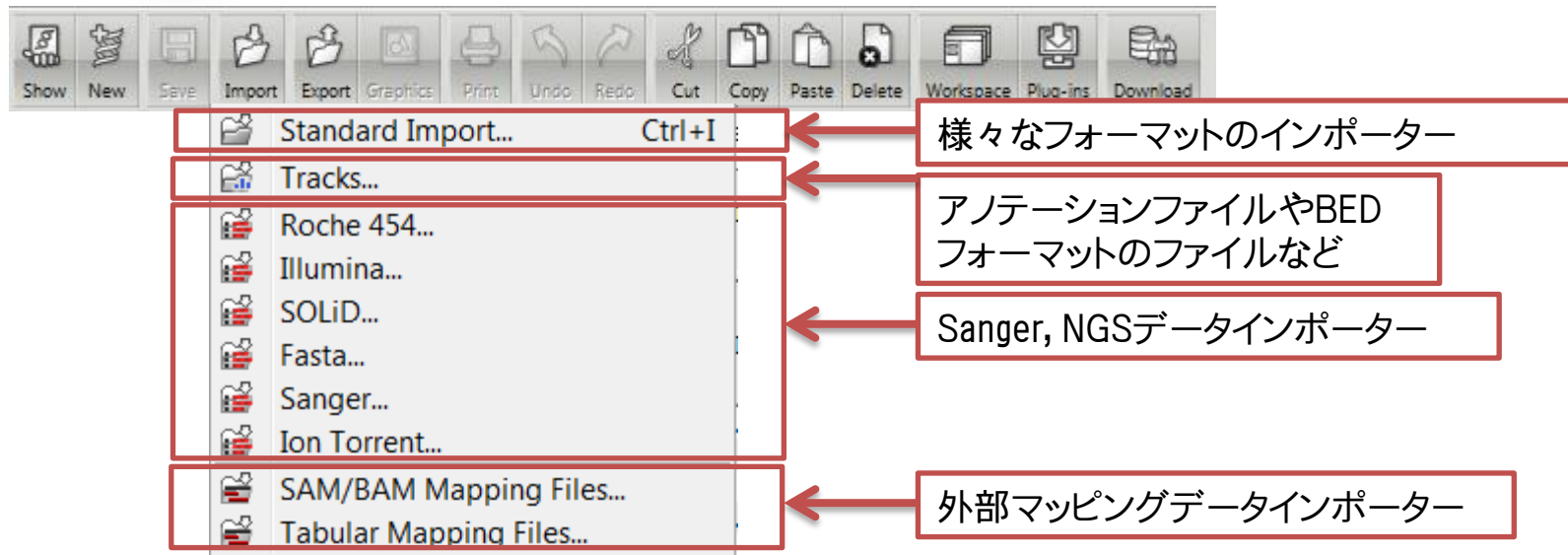


変異検出の流れ



リード・ゲノム・アノテーションインポート

データインポート

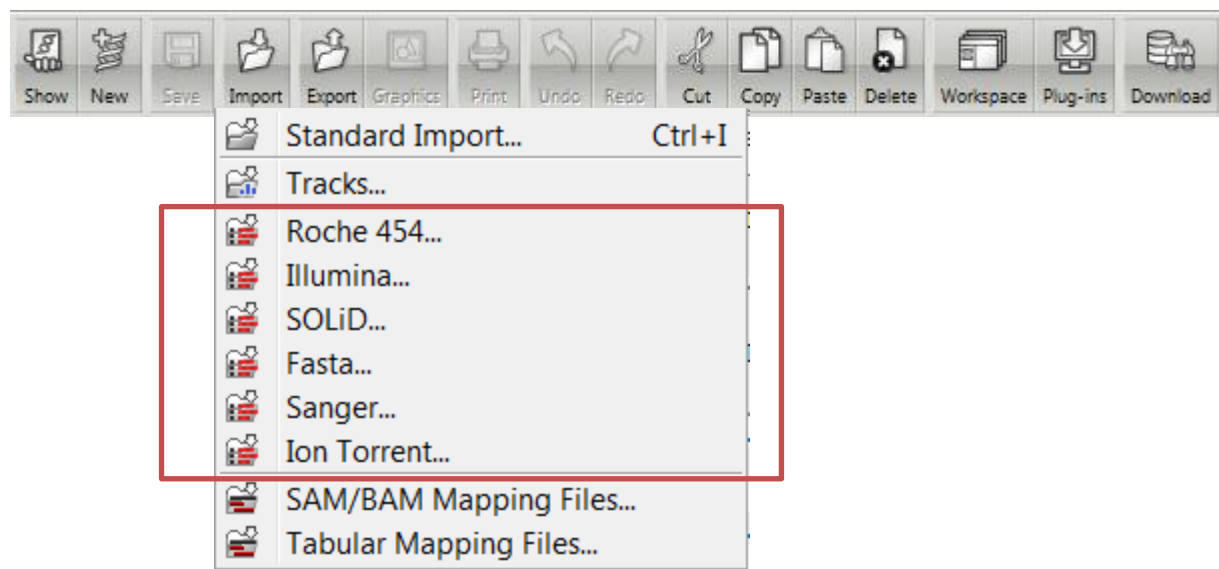


- Standard Importは、サンガーシーケンサー、次世代シーケンサー以外のファイルのインポートに利用します。



リードデータインポート

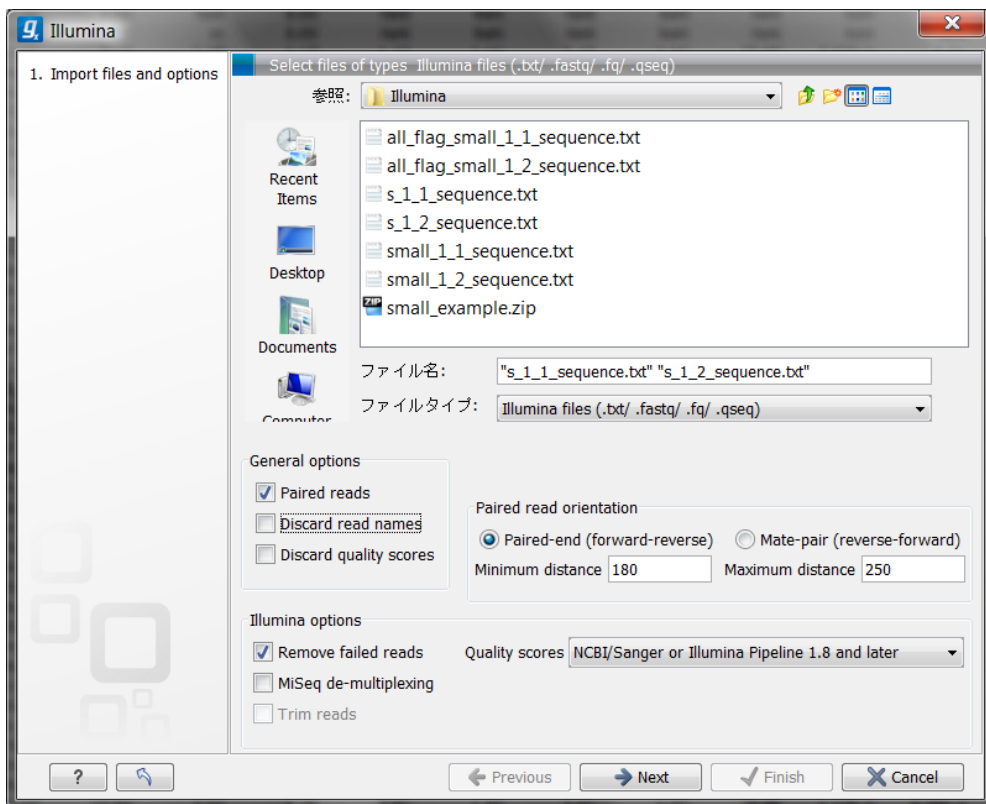
- Importからインポートしたいリードのシーケンサータイプを選択。





リードデータインポート

• Illuminaデータのインポート



General options

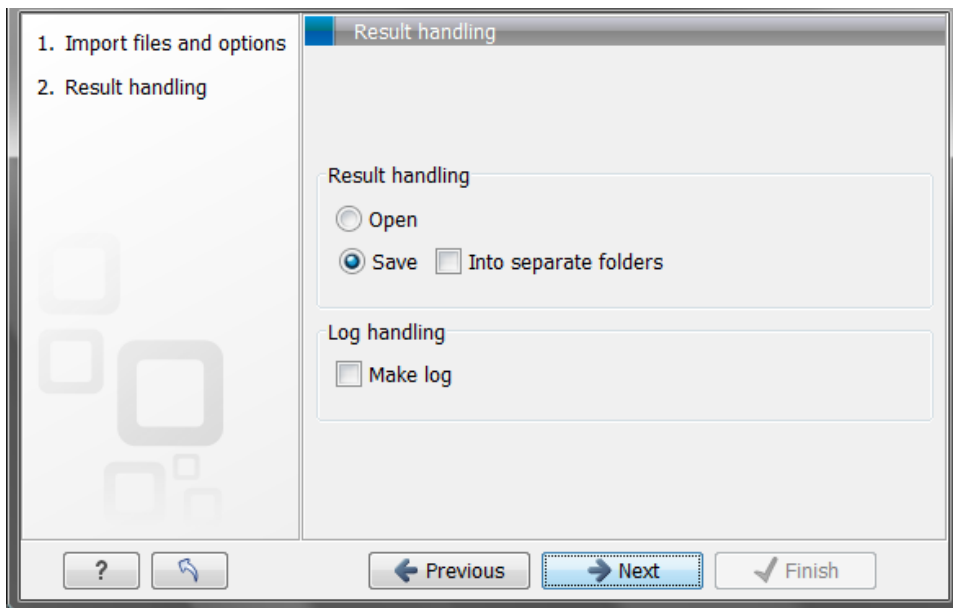
- Paired reads:ペアかどうか。
- Discard read names:リードについている名前を捨てるかどうか。デフォルトでは捨てるとなっておりますが、マッピング後、SAMにてExportした際など、リード名で確認したい場合があるため、最初は保存しておきましょう。
- Discard quality scores:Quality Scoreが必要ない場合はこのオプションにチェック。通常、インポート後にクオリティスコアが必要な事が多いです。
- Paired read orientation:ペアの距離と向きを指定。

Illumina options

- Remove failed reads:シーケンサーでfailとマークされたリードを除去するかどうか。
- MiSeq de-multiplexing:MultiplexingされたデータをDe-multiplexingするかどうか。
- Quality Score:使用するQuality Scoreのバージョンの選択。

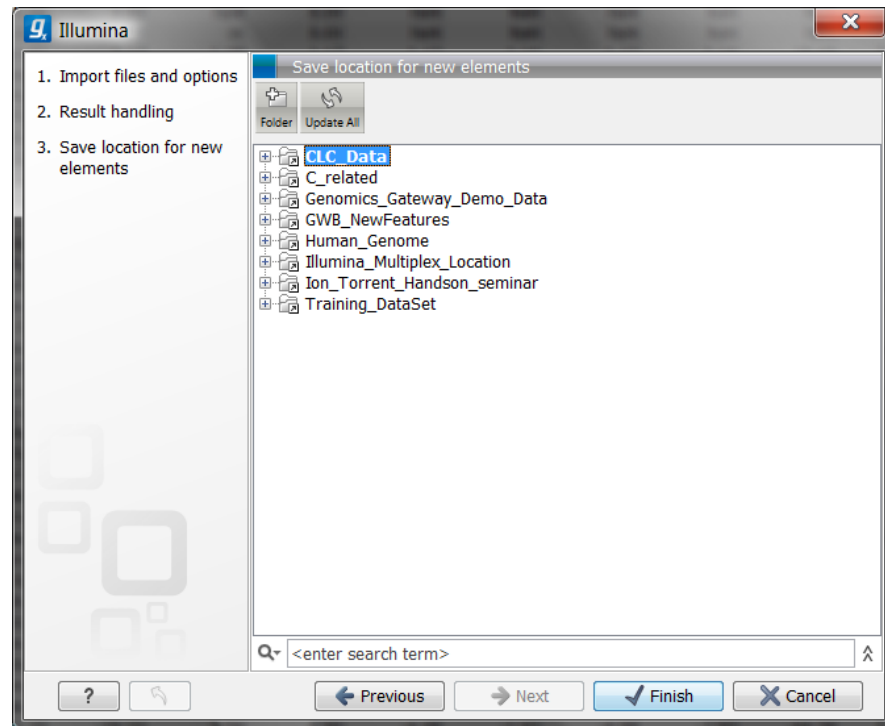


リードデータインポート



Result handling

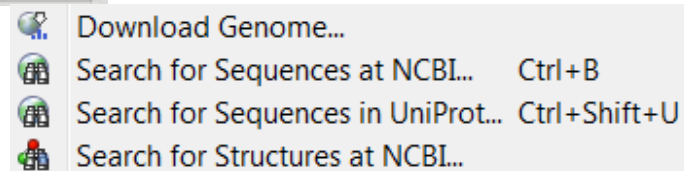
- データを開くか、保存の選択
- Into separate folders では、別々のフォルダへ保存するかどうかを選択できます。バッチ処理を行う際に便利です。



保存先の指定

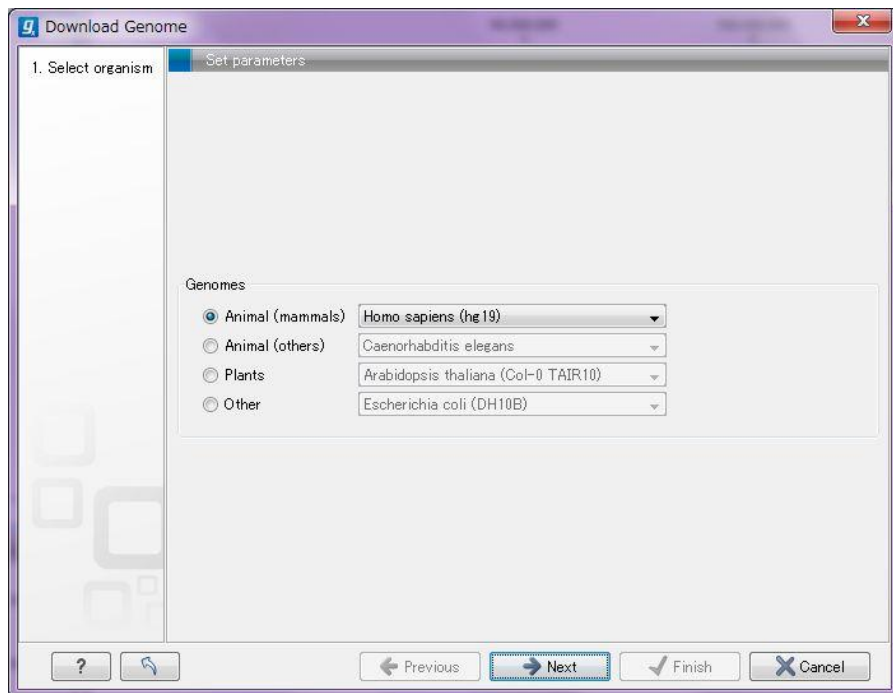
ゲノム・アノテーションインポート

- ゲノムはダウンロードアイコンより、生物種を指定してアノテーションと共にインポートすることが可能です。
- ゲノム配列とともに、アノテーションファイルをダウンロードすることも可能です。
- すでにGenomics Workbenchへ取り込んでいるゲノム配列について、アノテーションを付加することも可能です。

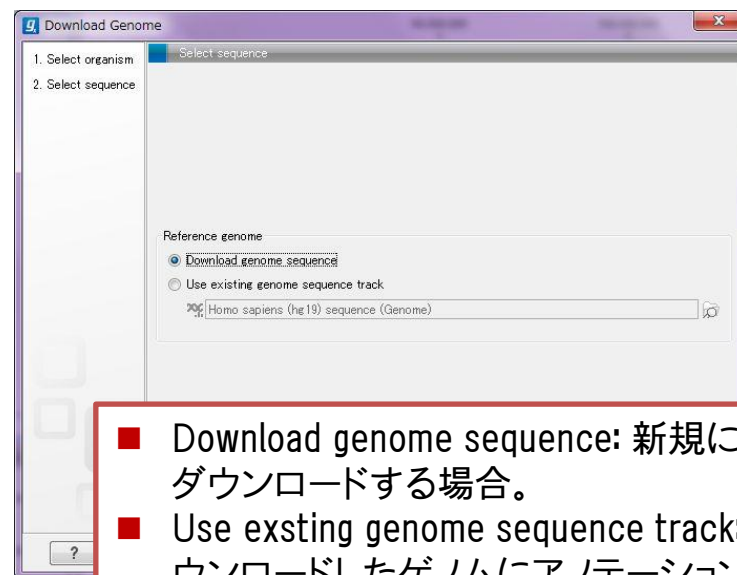




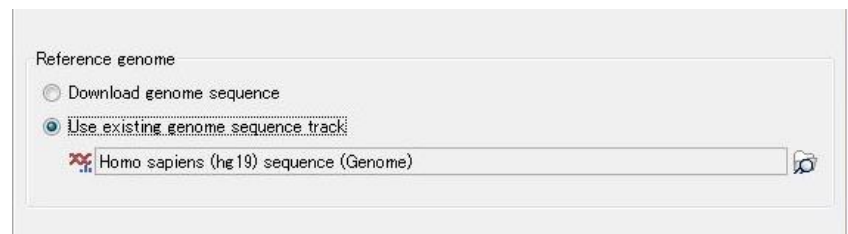
ゲノム・アノテーションインポート



- 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・アノテーションインポート

Download Genome

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

Select annotations

Download	Name	Provider	Size (in Mb)
<input type="checkbox"/>	Sequences	Ensembl	793.84
<input checked="" type="checkbox"/>	Gene annotation	Ensembl	20.35
<input type="checkbox"/>	Dbsnp (common) variants	UCSC	530.23
<input type="checkbox"/>	Dbsnp variants	UCSC	1473.00
<input type="checkbox"/>	COSMIC	SANGER	5.56
<input type="checkbox"/>	HapMap variants	Ensembl	362.98
<input type="checkbox"/>	1000Genomes variants	Ensembl	2294.97

Total download size 20.35 Mb

? ↶ ↷ 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。

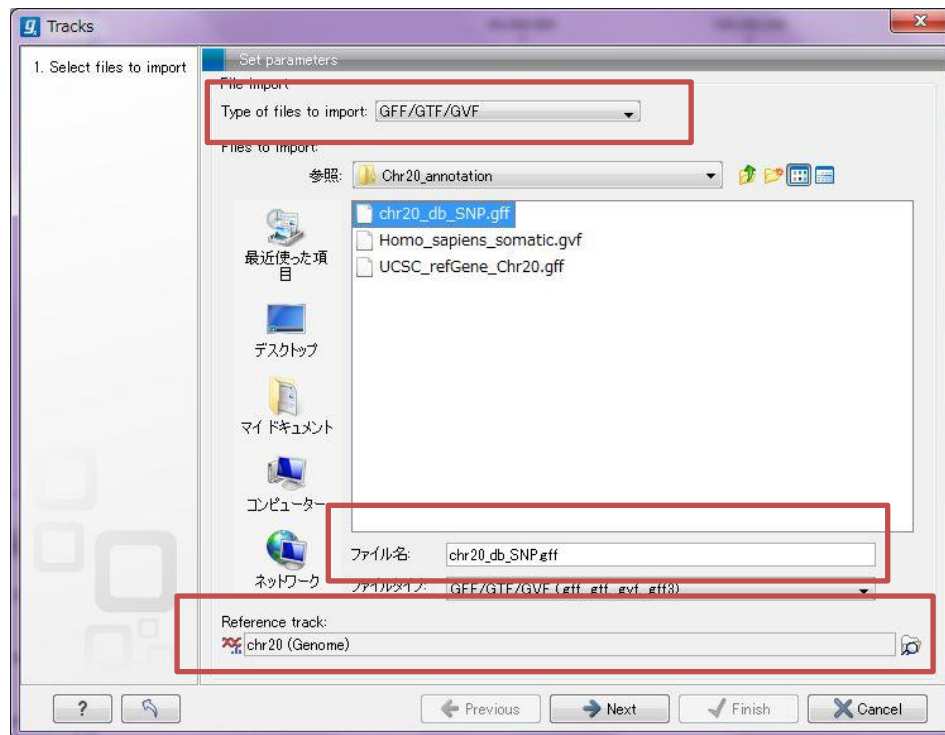
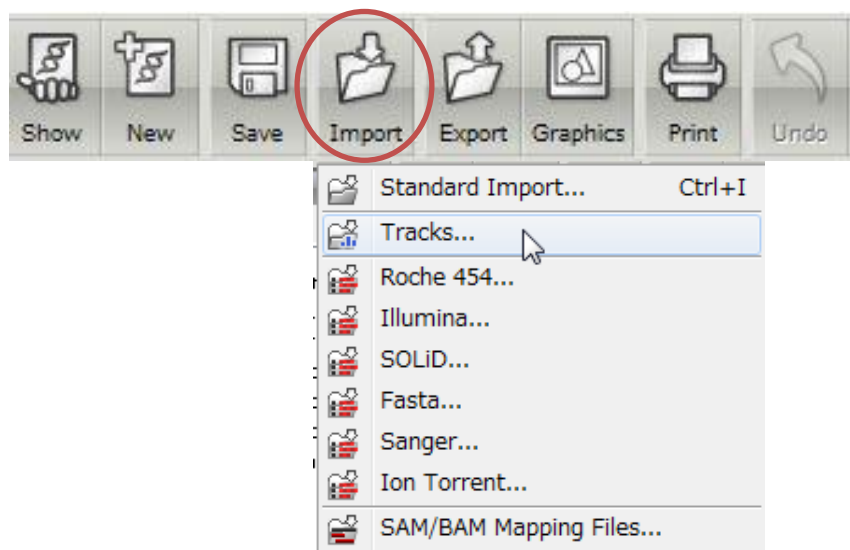


アノテーションインポート

- 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



アノテーションインポート



- Type of files to importを選択
- インポートするファイルを選択
- Reference Track を選択

クオリティチェック

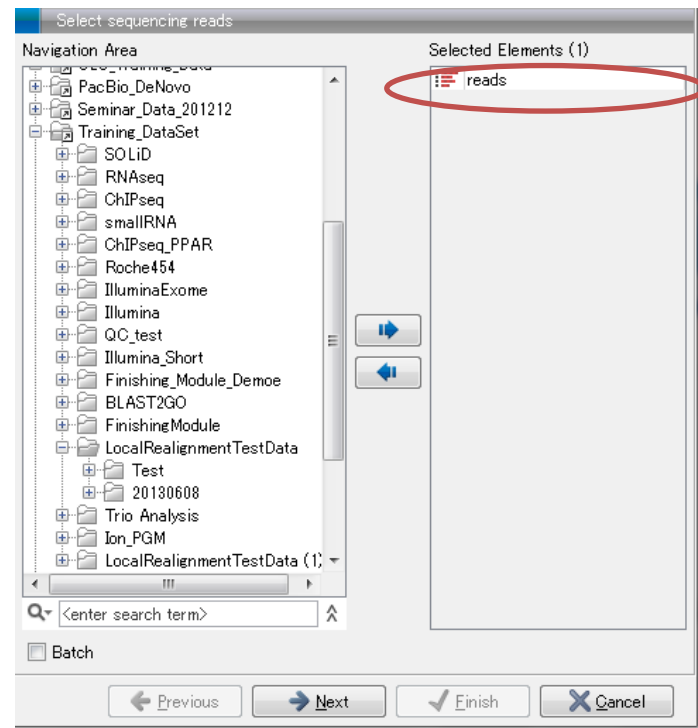
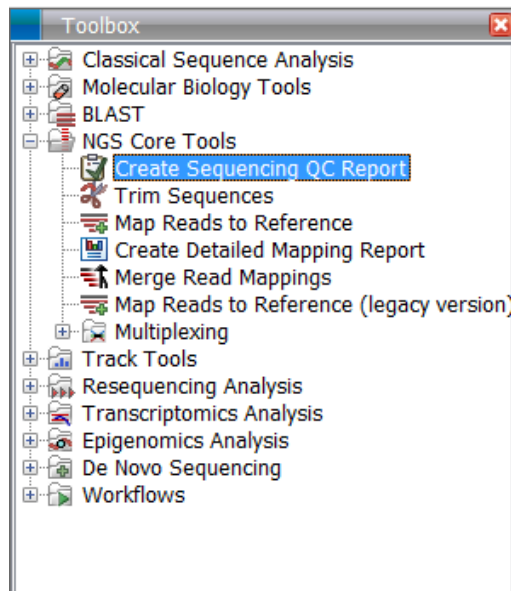
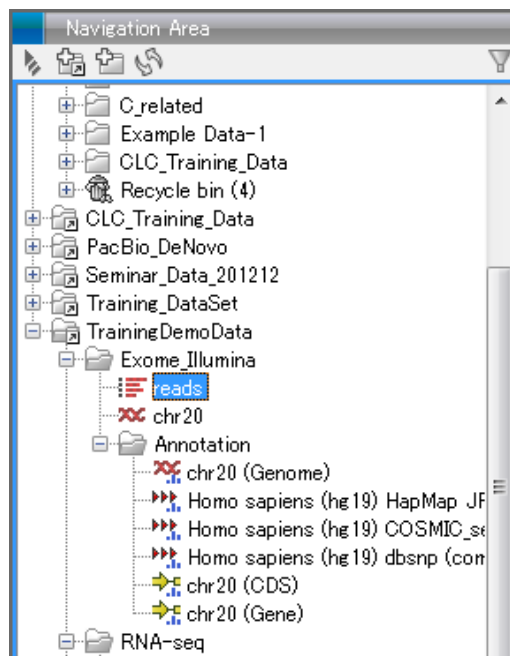


クオリティチェック流れ

- 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。
- PCR Duplicate の除去: Remove Duplicate Reads
 - フラグメント作成の過程でPCRが異常にかかってしまったものを補正。
- 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。
 - PCR Duplicateの除去は本日は行いませんが、行い方はWebの日本語マニュアルを参照してください。



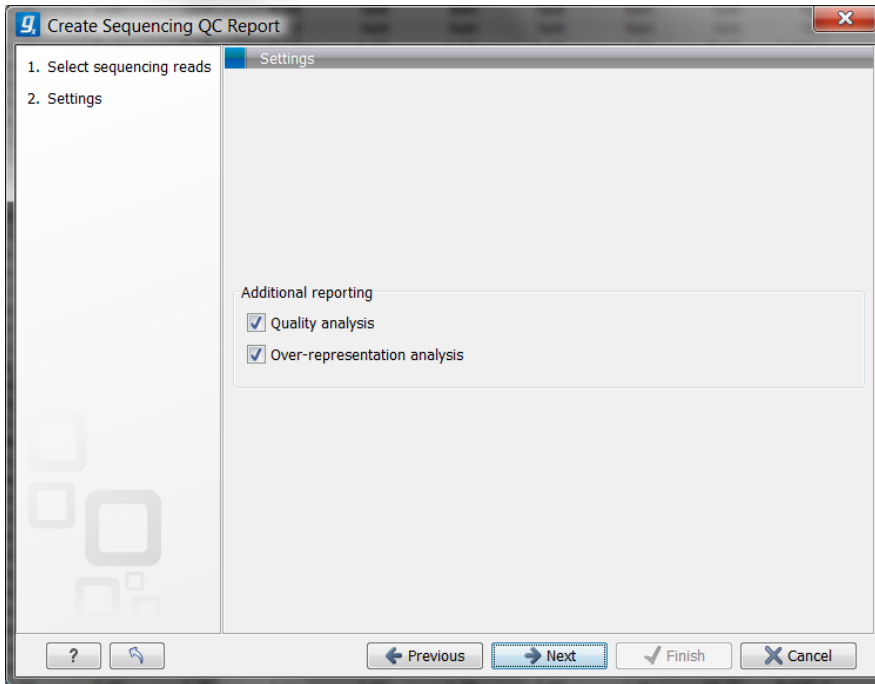
Create Sequencing QC Report



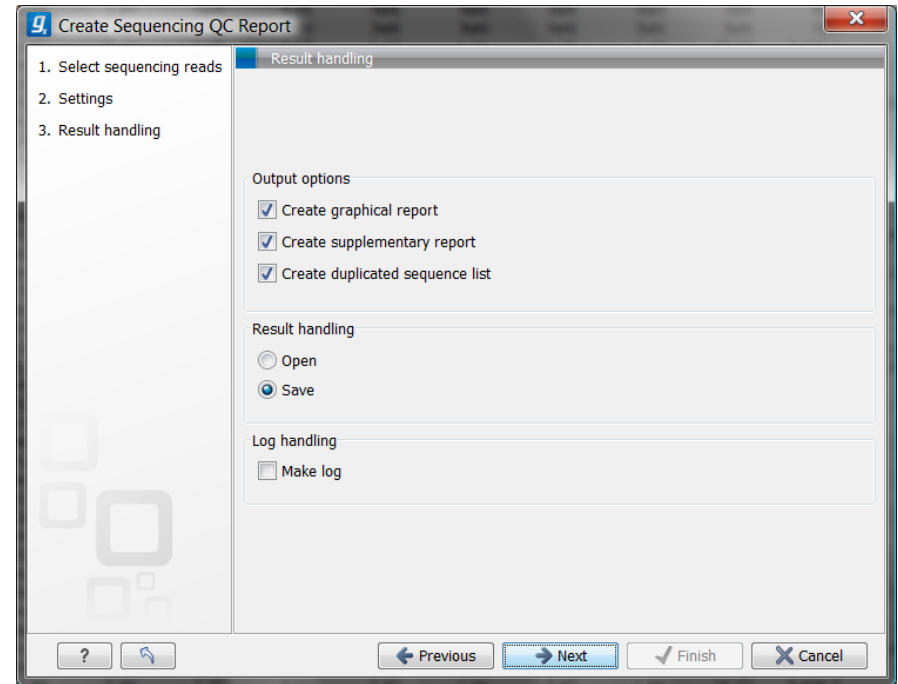
- Navigation Areaから使用するリードデータを選択。
- Toolboxから NGS Core Tools > Create Sequencing QC Report を選択、ダブルクリック。
- ウィザードが起動し、選択したデータが選ばれていることを確認。



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。

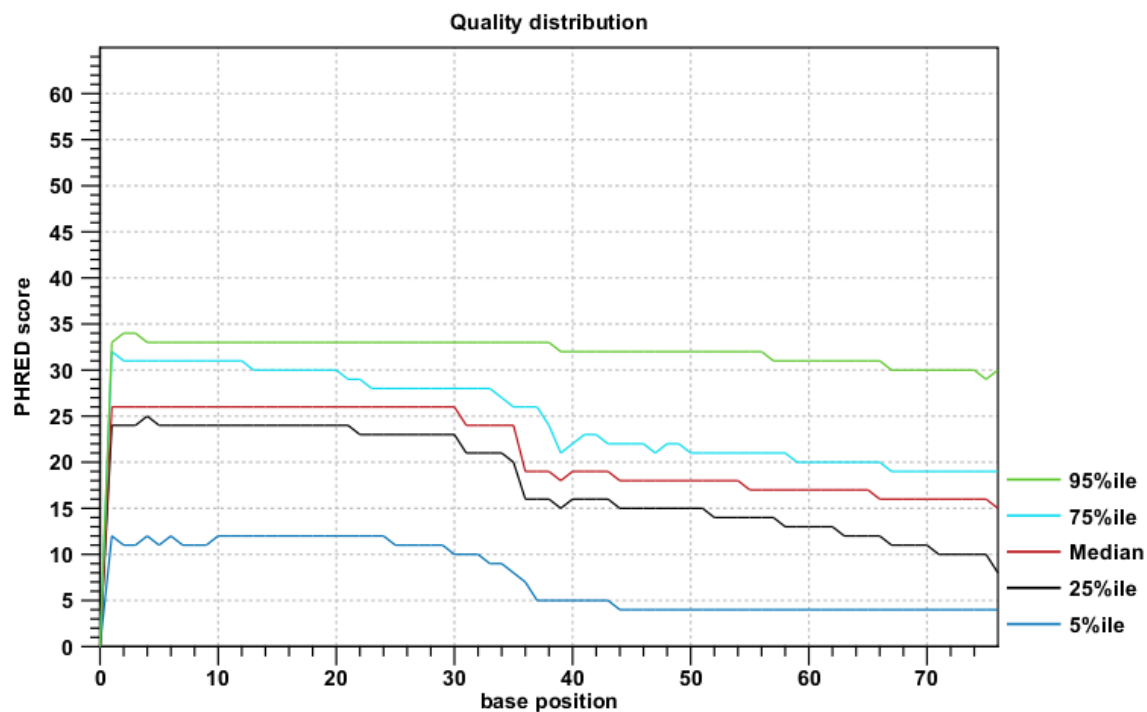


- Create graphical report: グラフィカルなレポート作成。
- Create supplementary report: 数値のレポート作成。
- Create duplicated sequence list: 重複のあった配列のリスト作成。

QCレポート 結果

 reads - graphical QC report

3.5 Quality distribution





クオリティチェック

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



クオリティトリミング原理

- 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
 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



クオリティトリミング原理

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

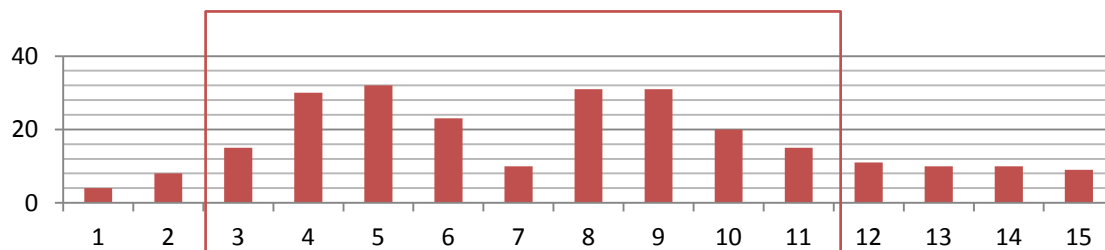


スタート点：
累積和が0より大きくなった塩基



終了点：
累積和が最大を示す塩基

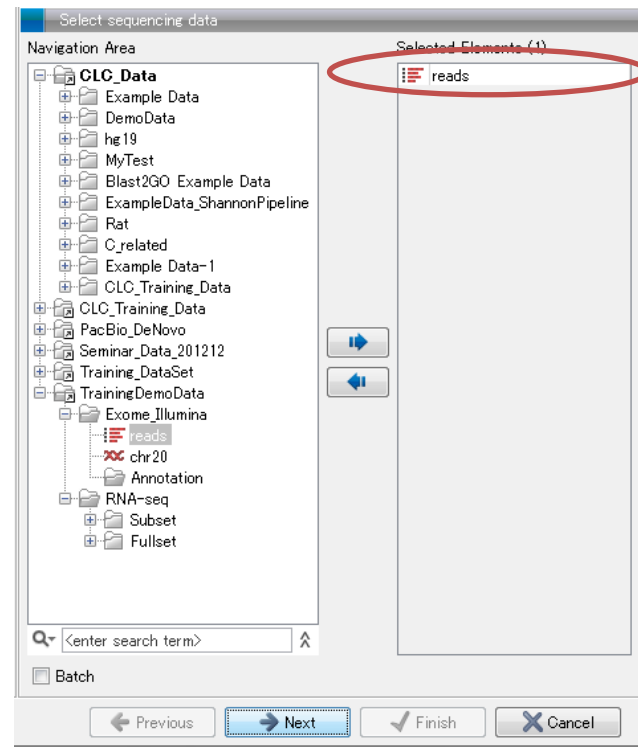
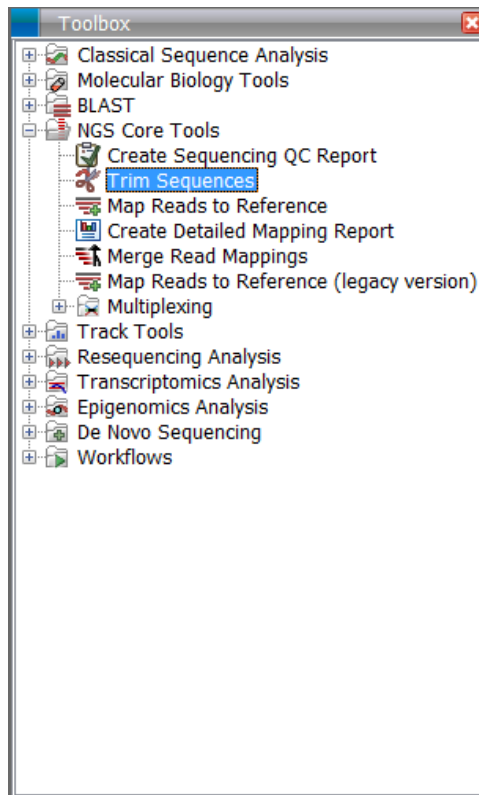
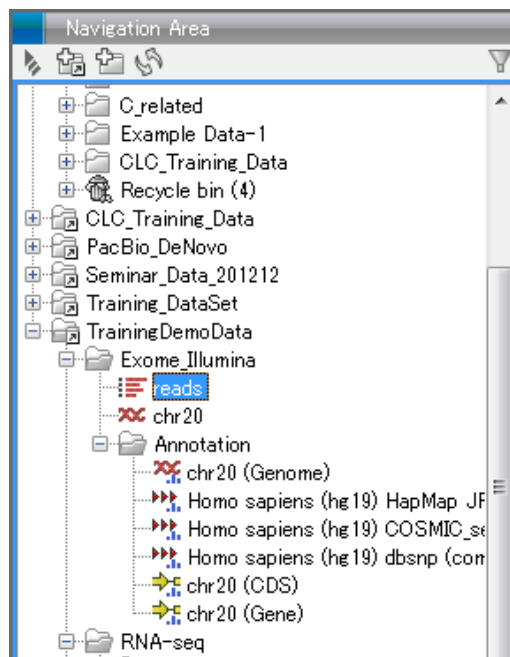
Phred score の棒グラフ



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。



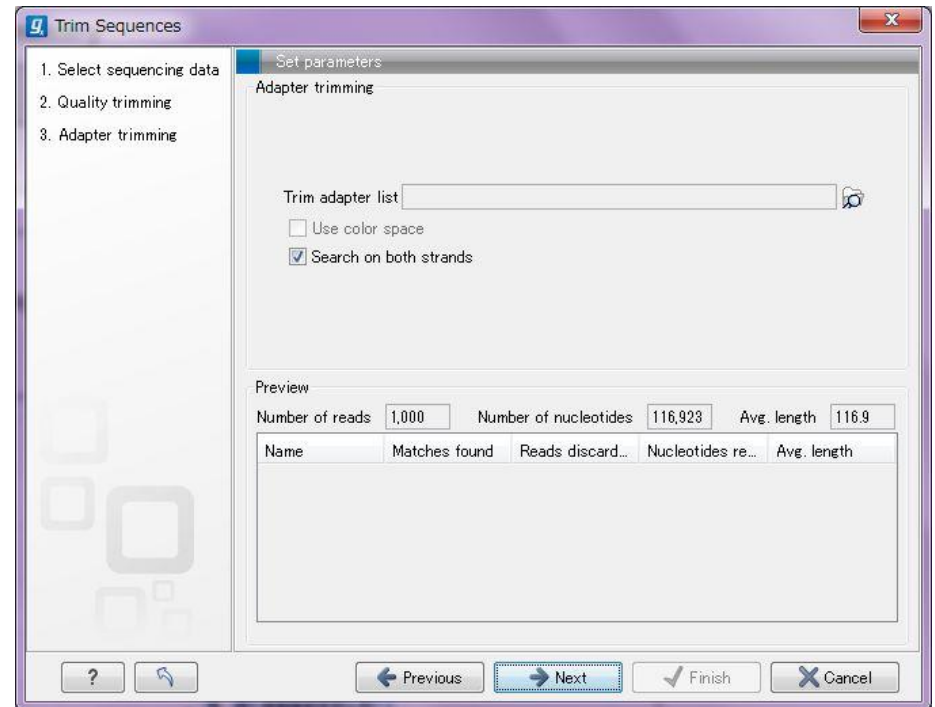
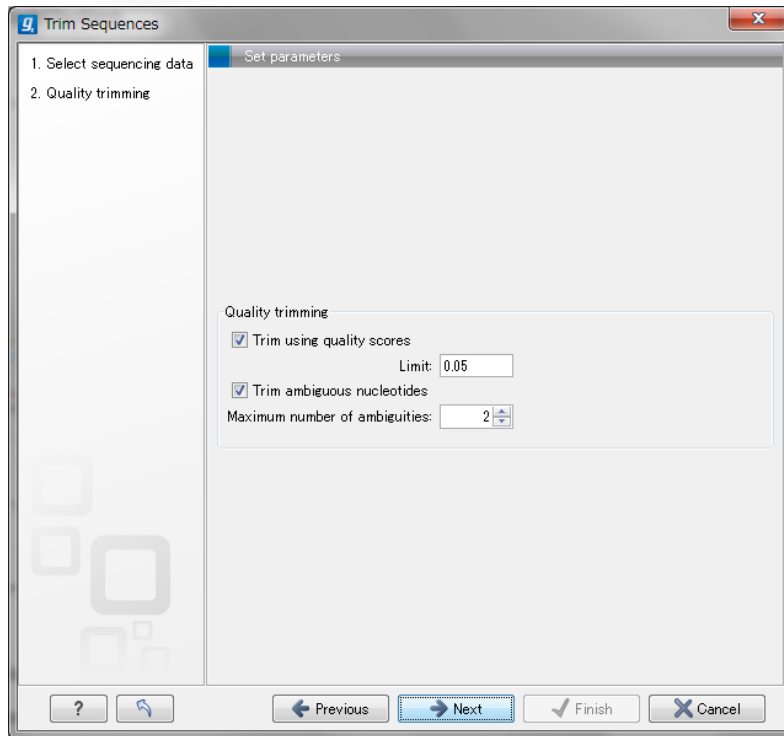
トリミング



- Navigation Areaから使用するリードデータを選択。
- Toolboxから NGS Core Tools > 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

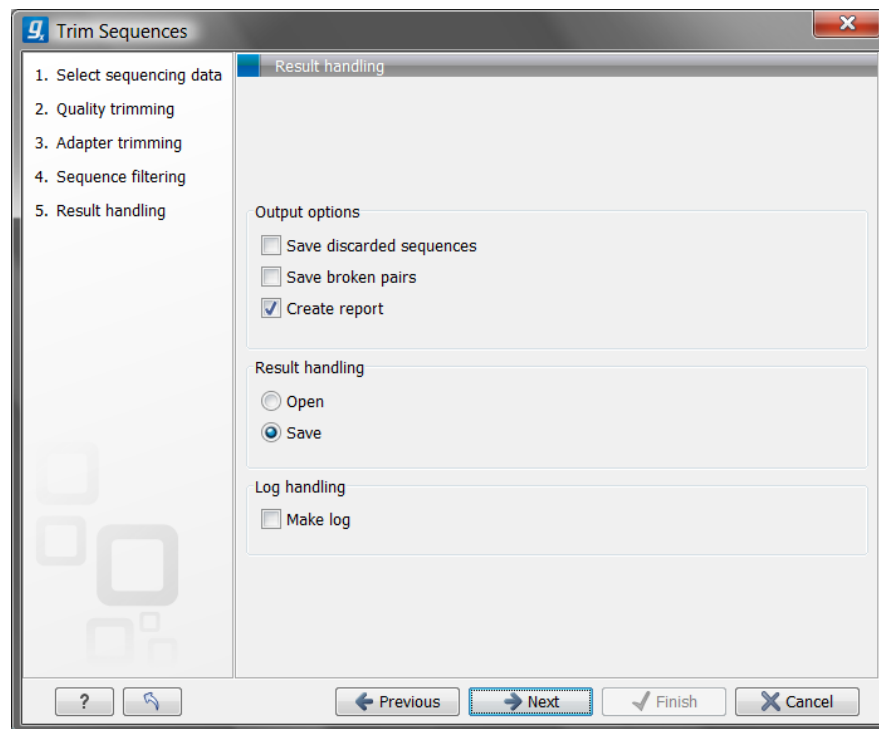
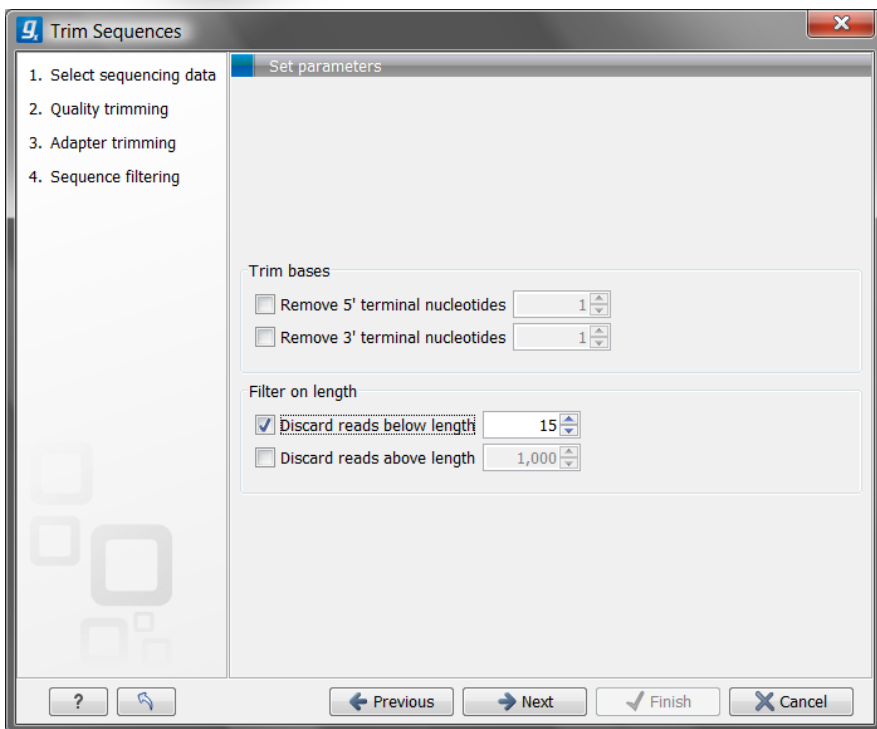


■ アダプタートリミング

- 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。



トリミング





- Trim bases:リード配列の5'末、3'末から指定数の塩基を除去
- Filter on length:リード配列の5'末、3'末から指定数の塩基を除去

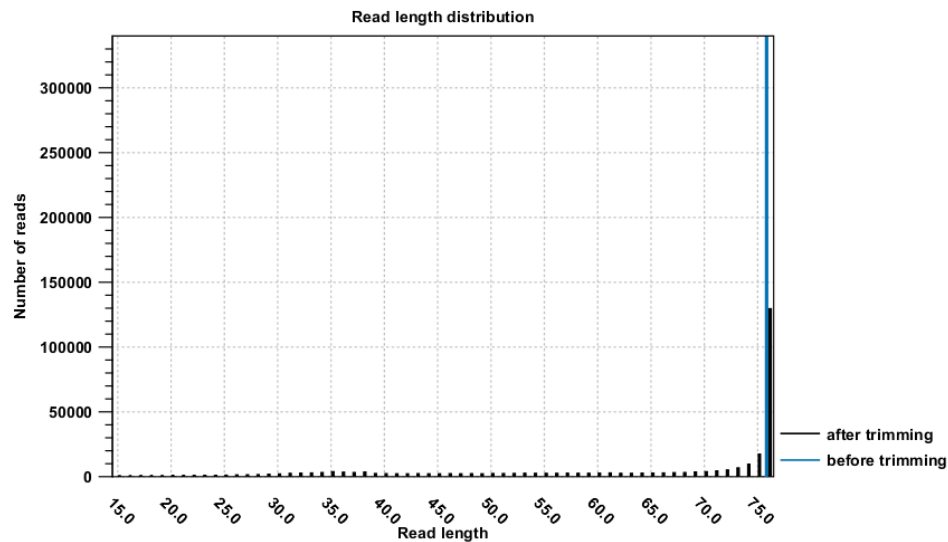
- Save discarded sequences: トリミングにより除去された配列の保存。
- Save broken pairs: ペアのリードでトリミングによりペアでなくなったリードを保存。
- Create report: レポートの作成。



トリミング結果

 reads trimmed
 reads report

2 Read length before / after trimming



■ 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。

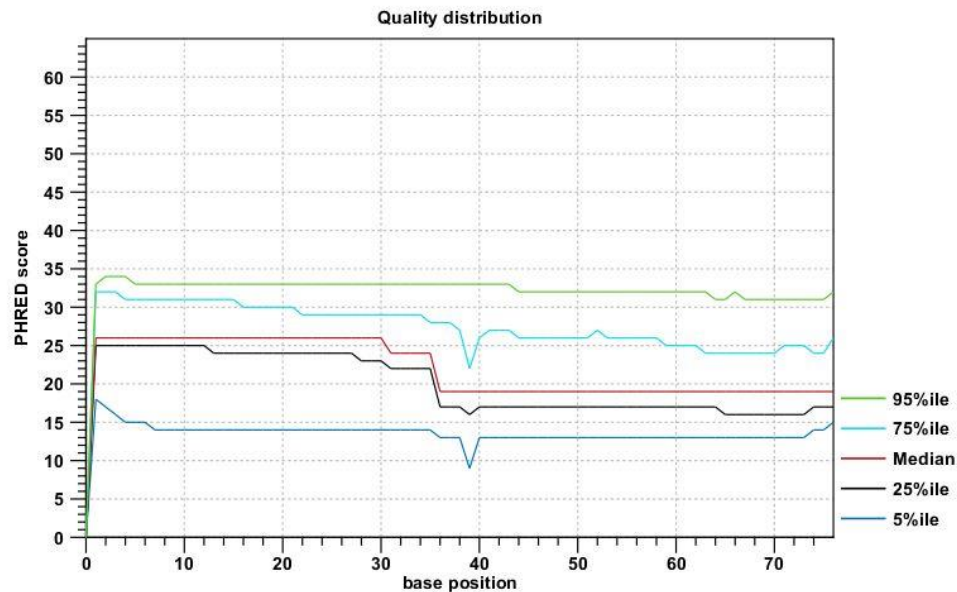
■ トリミング後は、トリムされたリードと、レポートを作成した場合は、そのレポートが作成されます。



QCレポート 再作成！

- トリミングされたリードを使って、QCレポートを再度作って、トリミング前と後を比較してみましょう。

3.5 Quality distribution

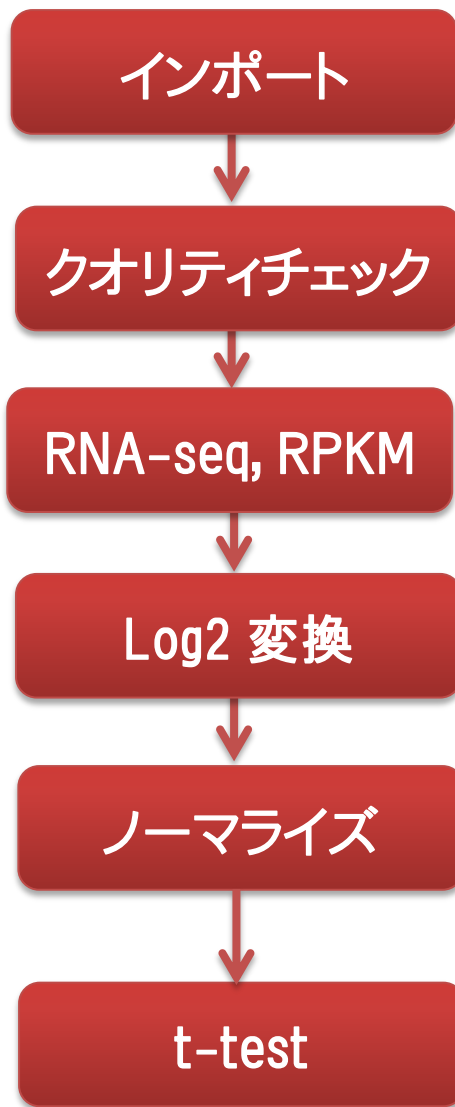


RNA-seq



- RNA-seq

RNA-seq解析フロー





RPKM

- RPKM: Reads Per Killobases per Million
 - 長さが異なるトランスクリプト、実験で使われたリードの総数による違いについて正規化するための方法。

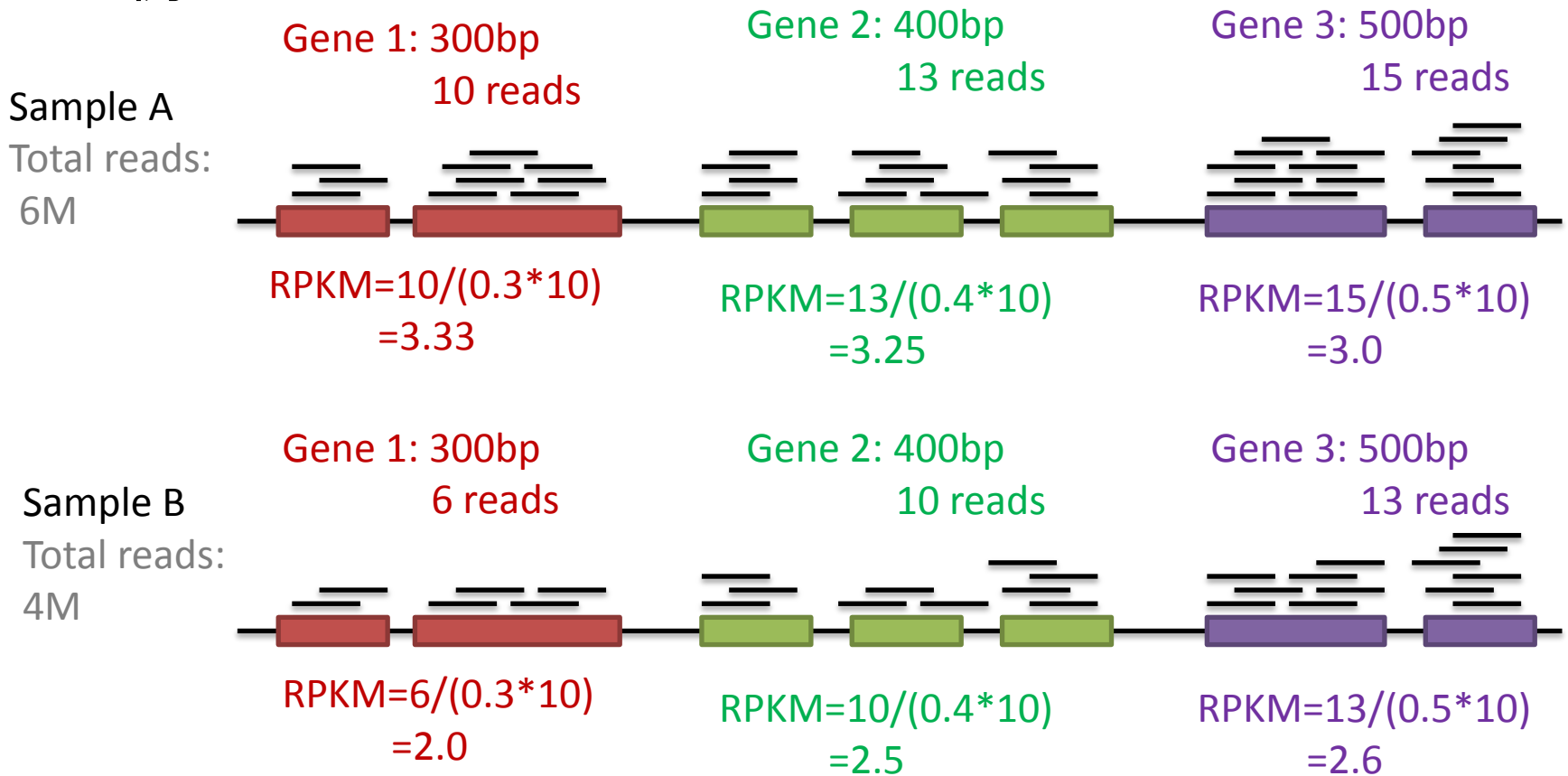
$$\text{RPKM} = \frac{C}{LN}$$

- C: マップされたリードの総数
- N: リードの総数(Million)
- L: トランスクリプトの長さ(kbase)



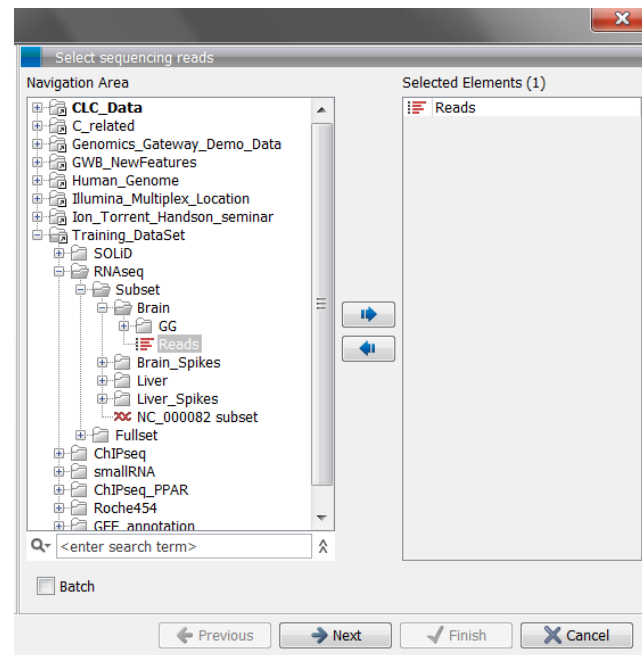
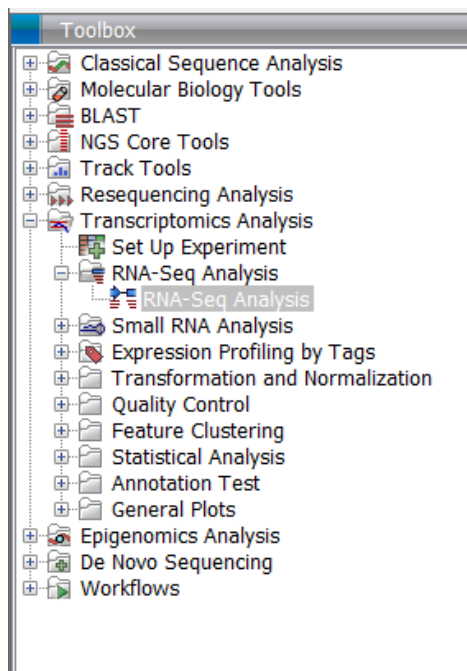
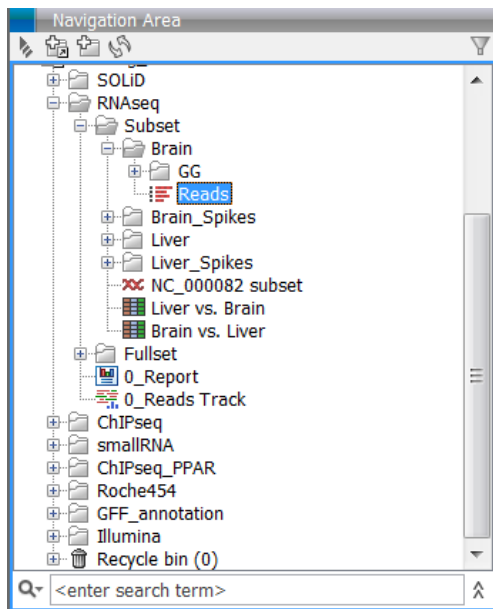
RPKM

• 例:





RNA-seq



- Navigation Areaから使用するリードデータを選択。
- Toolboxから Transcript Analysis > RNA-seq Analysis > RNA-Seq Analysis 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。



RNA-seq

RNA-Seq Analysis

1. Select sequencing reads

2. Set references

Set parameters

Reference

Use reference with annotations

Use reference without annotations

Extend annotated gene regions

Flanking upstream residues 0

Flanking downstream residues 0

Select annotated nucleotide sequences

Navigation Area

- CLC_Data
- C_related
- Genomics_Gateway_Demo_Data
- GWB_NewFeatures
- Human_Genome
- Illumina_Multiplex_Location
- Ion_Torrent_Handson_seminar
- Training_DataSet
- SOLID
- RNAseq
 - Subset
 - Brain
 - GG
 - Reads
 - Brain_Spikes
 - Liver
 - Liver_Spikes
 - NC_000082_subset
 - Fullset
 - ChIPseq
 - smallRNA
 - ChIPseq_PPAP
 - Roche454
 - GFF_annotation
 - Illumina

Selected Elements (1)

- NC_000082 subset

OK Cancel

Previous Next Finish Cancel



Reference

Use reference with annotations

Use reference without annotations

NC_000082 subset

Extend annotated gene regions

Flanking upstream residues 0

Flanking downstream residues 0

- アノテーション付のデータ、アノテーション無しのリファレンス、いずれかを選択。
- インポートしているゲノムのデータを選択。



RNA-seq

The screenshot shows the 'RNA-Seq Analysis' software window with the 'Set parameters' dialog box open. The dialog is divided into two main sections: 'Mapping settings' and 'Paired settings'. The 'Mapping settings' section includes four spinners: 'Maximum number of mismatches' (set to 2), 'Minimum length fraction' (set to 0.9), 'Minimum similarity fraction' (set to 0.8), and 'Maximum number of hits for a read' (set to 10). Below these are two checkboxes: 'Use color space' (unchecked) and 'Strand specific alignment' (checked). The 'Paired settings' section includes two text input fields: 'Minimum distance' (set to 180) and 'Maximum distance' (set to 250). Below these is a checkbox for 'Use 'include broken pairs' counting scheme' (unchecked). At the bottom of the dialog are buttons for '?', a refresh icon, 'Previous', 'Next', 'Finish', and 'Cancel'.

- Maximum number of mismatches: (Short read パラメータ) リード中に最大何個までのミスマッチを許容するか。
- Minimum length fraction: (Long read パラメータ) マッチする際に考慮するリードの長さの割合。
- Minimum similarity fraction: (Long read パラメータ) Minimum length fraction で指定した長さのうち、一致するべき割合。
- Maximum number of hits for a read: 1つのリードがマッチする最大の数。この数以上の箇所にマップされたリードは、マップされません。
- Use color space: カラースペースを使用する場合
- Strand specific alignment: センス鎖特異的にマップさせた場合のオプション
- Minimum distance: ペアの最小距離
- Maximum distance: ペアの最大距離
- Use 'include broken pairs' counting scheme: 指定した距離に納まらなかったリードもカウントしたい場合



RNA-seq

RNA-Seq Analysis

1. Select sequencing reads
2. Set references
3. Read mapping settings
4. Exon identification and discovery

Set parameters

Type of organism

Prokaryote
 Eukaryote

Exon discovery

Exon discovery

Required relative expression level 0.20

Minimum number of reads 10

Minimum length 50

? Refresh Previous Next Finish Cancel

- Exon discovery: 新規エクソンの探索を行いたい場合
- Required relative expression level: 新規エクソンとする場合に、その遺伝子の発現量のうち、どのぐらいの割合を持っている必要があるか。
- Minimum number of reads: 新規エクソンとする場合に最低限必要なリード数。
- Minimum length: 新規エクソンとする場合の最小の長さ。



RNA-seq

RNA-Seq Analysis

1. Select sequencing reads
2. Set references
3. Read mapping settings
4. Exon identification and discovery
5. Result handling

Result handling

Output options

- Create list of un-mapped sequences
- Create report
- Create fusion gene table

Minimum read count: 5

Expression value

Expression value: Genes: RPKM

Result handling

- Open
- Save

Log handling

- Make log

? ↻ Previous Next Finish Cancel

- Create list of un-mapped sequences: マップされなかったリードをリストとして回収するオプション
- Create report: レポート作成
- Create fusion gene table: Fusion gene の候補をリストで作成するかどうか。
- Minimum read count:(Pair-end オプション)作成する場合、Fusionとするための最小リードカウント。
- Expression value: デフォルトはRPKM。このほか、Total Exonなども選択可。後で変更も可能。



RNA-seq

Reads RNA-Seq x

Rows: 181 Filter:

Feature ID	Expressi...	Transcri...	Detected...	Exon len...	Unique g...	Total ge...	Unique e...	Total exo...	Ratio of ...	Unique e...	Total exo...	Unique in...	Total intr...	Exons	Putative ...	RPKM	Median C...	Chromos...	Chr
Dgcr6	6,471.68	1	1	1262	1215	1215	959	959	1.00	89	89	14	14	6	1	6,471.68	19.00	52954	
Prodh	3,159.62	1	1	2283	938	938	847	847	1.00	105	105	2	2	14	0	3,159.62	8.00	71820	
Rtn4r	6,671.36	1	1	1874	1547	1547	1468	1468	1.00	12	12	0	0	2	0	6,671.36	19.00	127800	
EG665975	52.96	1	1	1608	14	14	10	10	1.00	0	0	3	3	8	0	52.96	0.00	217973	
Zdhhc8	3,586.42	1	1	4868	2111	2111	2050	2050	1.00	144	144	8	8	11	0	3,586.42	9.00	220847	
Ranbp1	4,910.45	1	1	1025	619	619	591	591	1.00	73	73	3	3	6	0	4,910.45	13.00	239908	
Htf9c	2,164.31	3	3	2601	679	679	661	661	1.00	57	57	11	11	14	0	2,164.31	6.00	248977	
Dgcr8	1,013.67	1	1	4226	539	539	540	503	1.00	40	40	1	1	14	0	1,013.67	3.00	254058	
D16H22S...	8,163.27	1	1	1471	1464	1465	1410	1410	1.00	168	168	2	2	8	0	8,163.27	22.00	300919	
LOC10004...	276.96	1	1	369	9	9	13	9	0.75	0	0	0	0	3	0	276.96	0.00	348262	
Arvcf	1,611.57	1	1	4624	902	938	844	875	0.96	81	83	1	1	19	0	1,611.57	4.00	348368	
Comt	11,591.06	1	1	1252	1736	1789	1651	1704	0.97	109	109	2	2	5	0	11,591.06	31.00	407637	
Txnrd2	1,589.56	1	1	1902	473	484	344	355	0.97	77	77	11	11	18	0	1,589.56	4.00	426511	
Gnb1l	219.33	2	1	3650	211	211	94	94	1.00	13	13	0	0	9	0	219.33	0.00	498862	
Tbx1	249.21	1	1	1777	59	59	52	52	1.00	2	2	1	1	7	0	249.21	0.00	581807	
Gp1bb	1,642.02	2	1	2085	318	402	318	402	0.79	0	0	0	0	3	0	1,642.02	4.00	620413	
Sept5	64,093.18	1	1	2130	13531	17547	12014	16030	0.75	1915	1915	29	29	12	0	64,093.18	188.00	621905	
LOC622795	0.00	0	0	0	17	17	0	0	NaN	0	0	0	0	0	0	0.00	0.00	681356	
Cldn5	8,630.87	1	1	1414	1433	1433	1433	1433	1.00	0	0	0	0	1	0	8,630.87	24.00	776941	
Cdc45l	191.20	1	1	2138	55	55	48	48	1.00	10	10	0	0	20	0	191.20	0.00	780546	
Ufd1l	2,505.35	1	1	1958	619	619	576	576	1.00	69	69	0	0	12	0	2,505.35	7.00	812420	
Z510002D...	871.00	1	1	1232	160	160	126	126	1.00	7	7	4	4	3	0	871.00	2.00	836674	
Mrp140	3,104.12	1	1	856	320	320	312	312	1.00	19	19	2	2	4	0	3,104.12	8.00	872311	
Hira	2,238.99	1	1	4534	1299	1300	1192	1192	1.00	143	143	5	5	25	0	2,238.99	6.00	876844	
Igf	0.00	0	0	0	27	27	0	0	NaN	0	0	0	0	0	0	0.00	0.00	1026952	1
Igf-C1	0.00	0	0	0	0	0	0	0	NaN	0	0	0	0	0	0	0.00	0.00	1061846	1
Igf-J1	0.00	0	0	0	0	0	0	0	NaN	0	0	0	0	0	0	0.00	0.00	1063319	1
Igf-C3	0.00	0	0	0	0	0	0	0	NaN	0	0	0	0	0	0	0.00	0.00	1065459	1
Igf-13p	0.00	0	0	0	0	0	0	0	NaN	0	0	0	0	0	0	0.00	0.00	1066465	1
Igf-13	0.00	0	0	0	0	0	0	0	NaN	0	0	0	0	0	0	0.00	0.00	1067135	1
Igf-V1	0.00	0	0	0	0	0	0	0	NaN	0	0	0	0	0	0	0.00	0.00	1085111	1
Igf-C4	0.00	0	0	0	0	0	0	0	NaN	0	0	0	0	0	0	0.00	0.00	1195093	1
Igf-14	0.00	0	0	0	0	0	0	0	NaN	0	0	0	0	0	0	0.00	0.00	1196589	1
Igf-C2	0.00	0	0	0	0	1	0	0	NaN	0	0	0	0	0	0	0.00	0.00	1198630	1
Igf-J2	0.00	0	0	0	0	0	0	0	NaN	0	0	0	0	0	0	0.00	0.00	1200292	1
Igf-V3	0.00	0	0	0	0	0	0	0	NaN	0	0	0	0	0	0	0.00	0.00	1241302	1
Igf-V2	0.00	0	0	0	0	0	0	0	NaN	0	0	0	0	0	0	0.00	0.00	1260497	1
Olfrl64	0.00	1	0	948	0	0	0	0	NaN	0	0	0	0	1	0	0.00	0.00	1285888	1

Gene-Level Expression Settings

Column width: Manual

Show column:

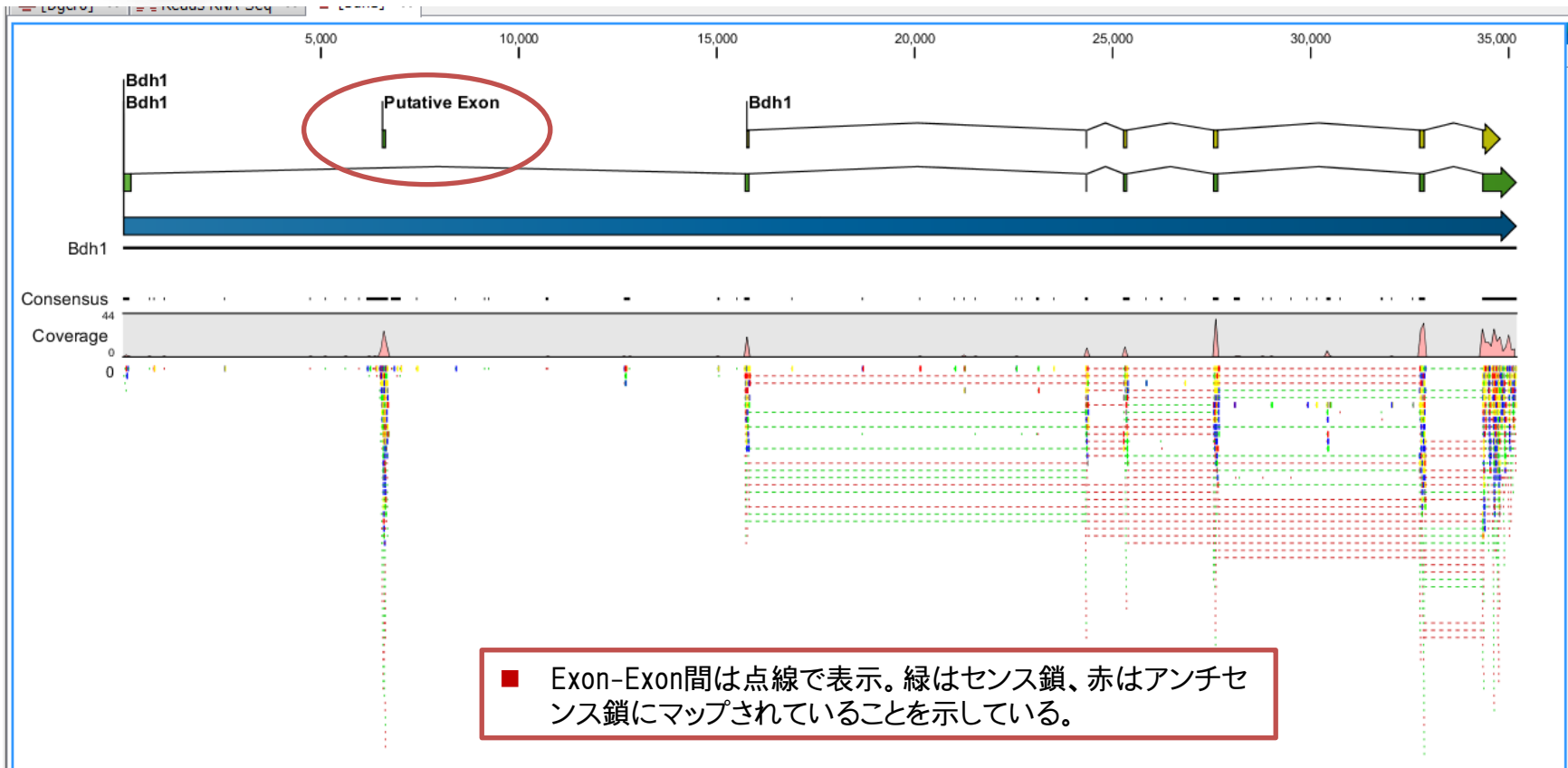
- Feature ID
- Expression values
- Transcripts annotated
- Detected transcripts
- Exon length
- Unique gene reads
- Total gene reads
- Unique exon reads
- Total exon reads
- Ratio of unique to total (exon reads)
- Unique exon-exon reads
- Total exon-exon reads
- Unique intron-exon reads
- Total intron-exon reads
- Exons
- Putative exons
- RPKM
- Median coverage
- Chromosome region start
- Chromosome region end

Select All

Deselect All



RNA-seq



Expression Analysis



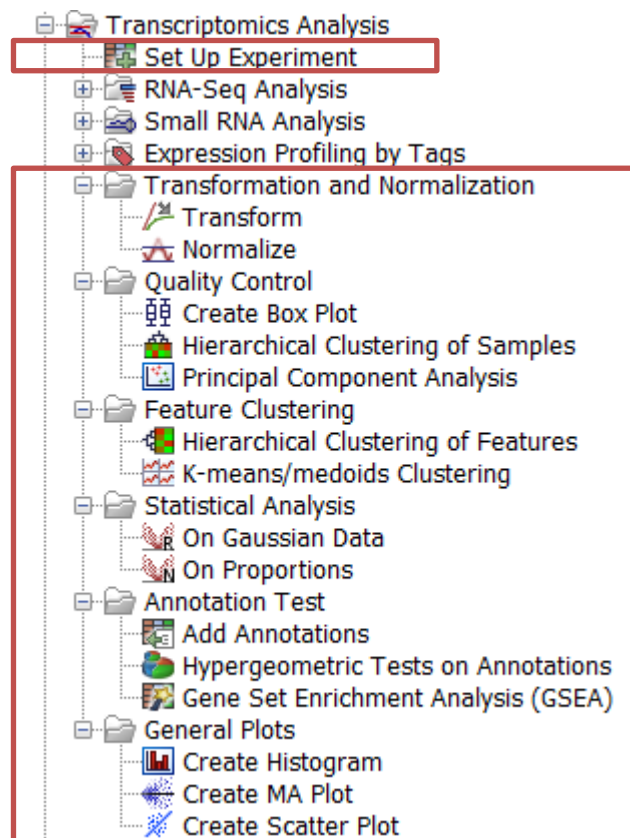
発現解析

- RNA-seqの結果を使から、「BrainとLiverにおいて有意に発現している遺伝子は何か」といったことを調べます。
- 群間の比較を行う場合、RPKMの値をそのまま使う方法と、遺伝子に張り付いたリードの数をそのまま使う方法の2種類があります。

		群	群内のレプリケート
Gaussian Test	T-test	2群	必須
	ANOVA	3群以上	必須
Proportional Test	Kal's test	2群	不要
	Baggerley's test	2群	必須

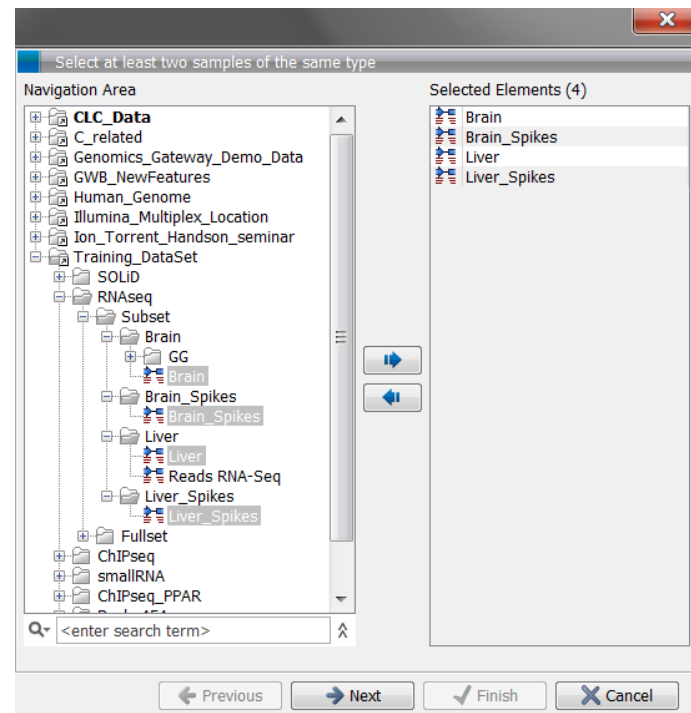
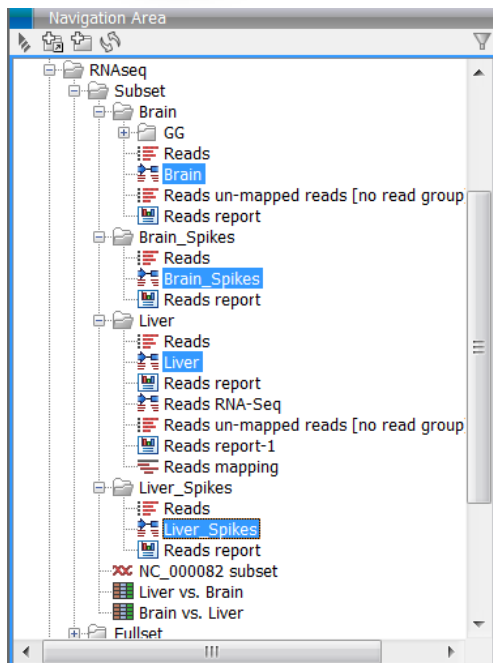
Expression Analysis

- RNA-seqのデータはMicroarrayのように発現差の解析を行うことが可能です。
- そのためには、まずRNA-seqのデータをExperimentという形へ変更し、その後、発現解析ツールを使って解析を行います。





Expression Analysis



- Navigation Areaから使用するRNA-seqデータを選択。
- Toolboxから Set Up Experiment 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。



Expression Analysis

9 Set Up Experiment

1. Select at least two samples of the same type

2. Define experiment type

Define experiment type

Experiment

Two-group comparison

- Unpaired
- Paired

Multi-group comparison

- Unpaired
- Paired

Number of groups: 2

Expression values

Use existing expression values from samples

Set new expression value

Value to use in experiment: Genes: Unique exon reads

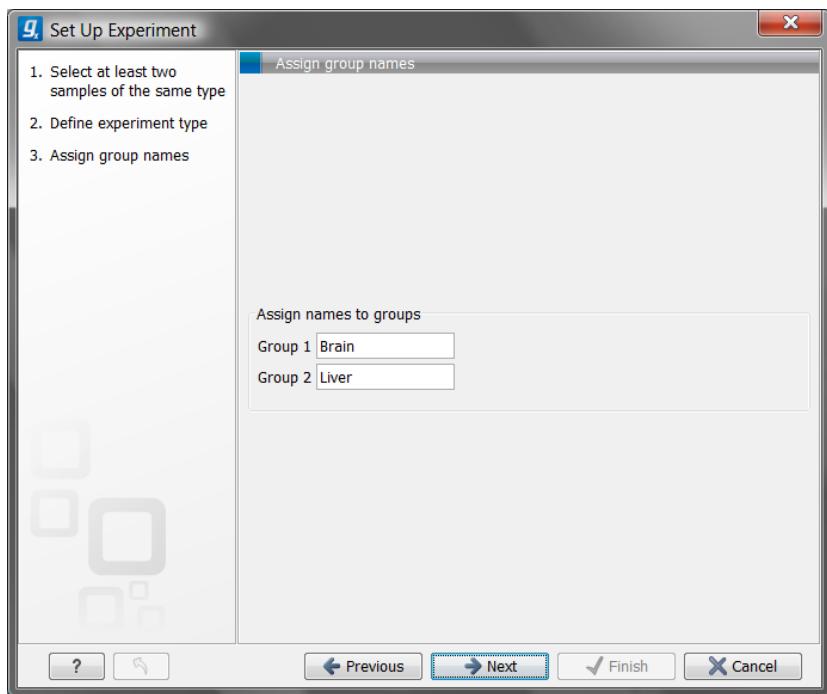
? ↶ ↷ ✓ Finish ✕ Cancel

- Two-group comparison: 2群比較
 - Unpaired/Paired:2つの群のサンプルに対応があるかどうか。(同じ固体で違う条件など)
- Multi-group comparison:多群比較
- Use existing expression values from samples: RNA-seqで指定した発現量をそのままつかう場合。
- Set new expression value: 別の発現量を使う場合。

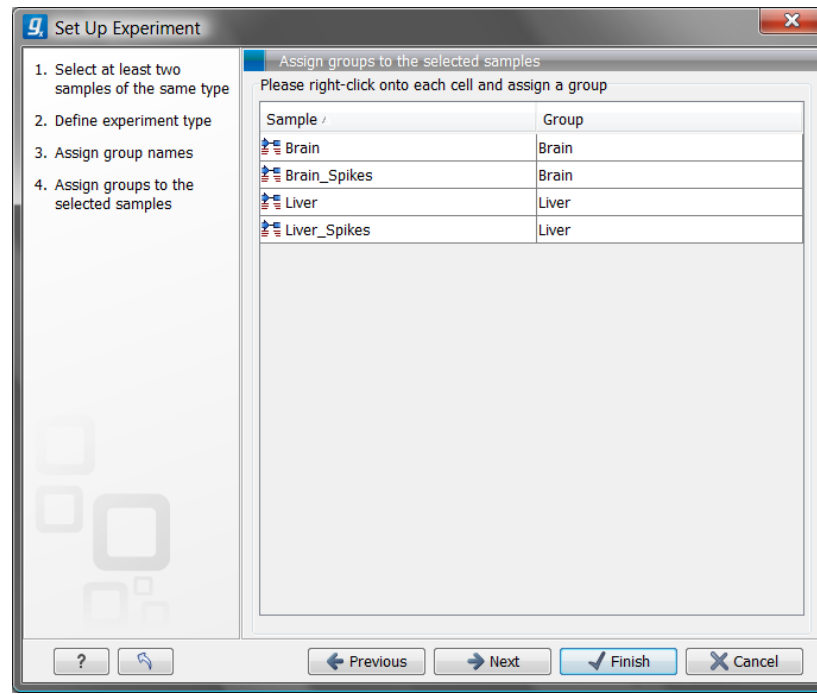
- Experimentを作成する際は、ひとまず何かの検定を行うこととなります。



Expression Analysis



■ グループにつける名前を入力



■ RNA-seqのデータをグループに割り当てる

Expression Analysis

Brain vs. Liver x

Rows: 181 Filter:

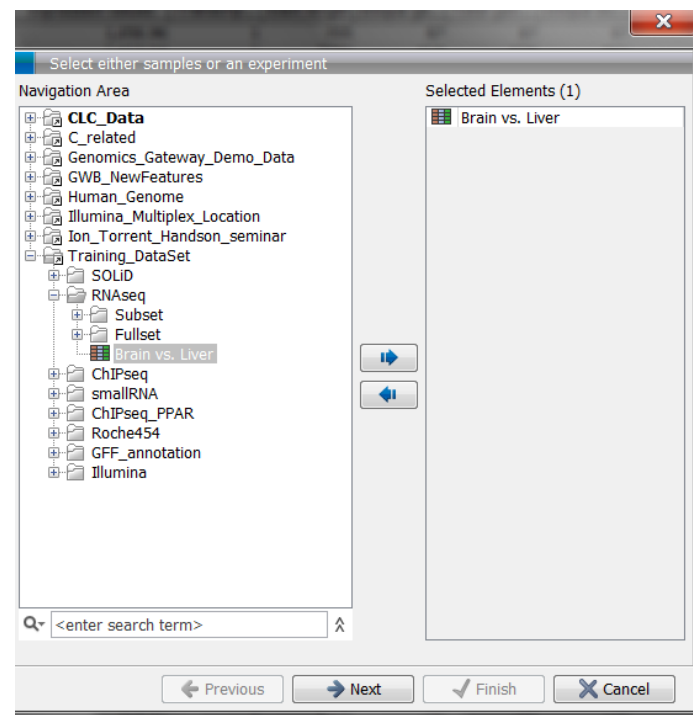
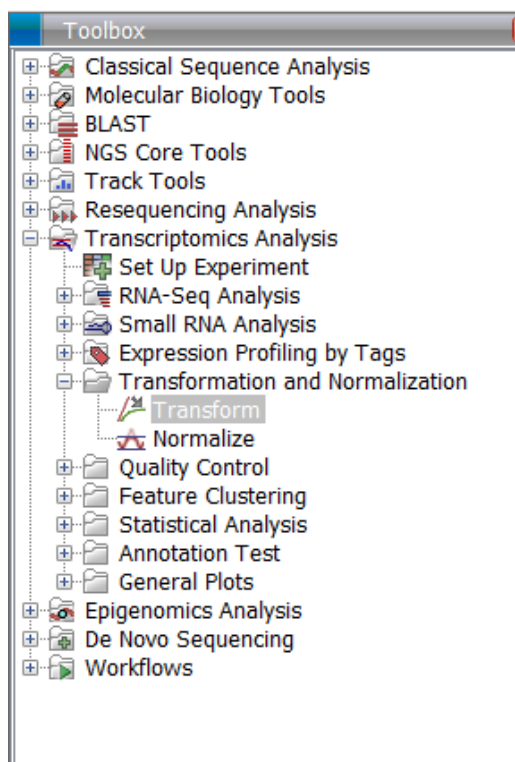
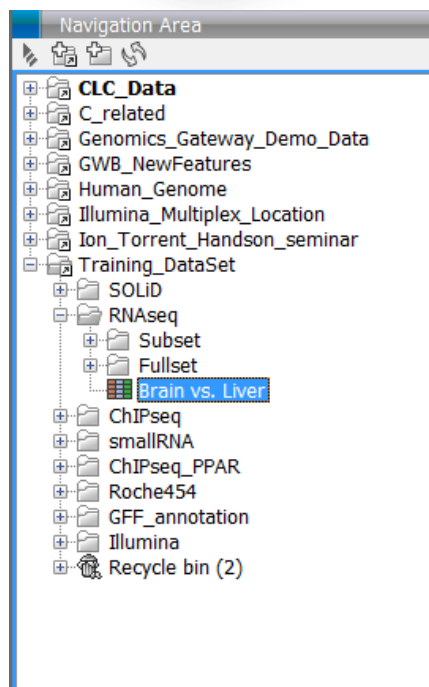
Feature ID	Experiment				Brain											
	Range (or...)	IQR (origi...)	Differenc...	Fold Chan...	Expressio...	Transcrip...	Exon length	Unique ge...	Total gen...	Unique ex...	Total exo...	Unique ex...	Unique int...	Total intr...	Exons	Pi
1110054...	599.49	500.74	-509.47	-2.02	1,050.96	1	705	87	87	87	87	0	0	0	0	1
1600021P...	1,422.81	1,421.10	-1,391.08	-6.52	1,612.04	1	3080	619	620	583	583	6	0	0	2	
1700007E...	0.00	0.00	0.00	1.00	0.00	0	0	0	0	0	0	0	0	0	0	
1700025H...	5.94	5.94	2.97	∞	0.00	1	547	0	0	0	0	0	0	0	2	
1810009K...	337.32	335.49	313.41	2.17	268.30	1	984	35	35	31	31	2	2	0	3	
2310042E...	18.35	16.42	-14.89	-16.47	13.35	1	1276	0	2	0	2	0	0	0	1	
2510002D...	494.02	466.47	-455.21	-2.16	871.00	1	1232	160	160	126	126	7	7	4	3	
2510009E...	1,402.24	1,401.64	-1,377.14	-22.15	1,417.46	1	5113	885	885	851	851	3	3	1	2	
2900046G...	15,812.44	15,806.41	-15,335.32	-1,821.61	15,817.85	1	2271	4275	4277	4216	4218	286	286	3	6	
5430420C...	4.32	4.32	-2.16	-∞	0.00	1	1084	28	28	0	0	0	0	0	6	
6430590L...	0.00	0.00	0.00	1.00	0.00	0	0	52	52	0	0	0	0	0	0	
9030404E...	0.00	0.00	0.00	1.00	0.00	1	488	8	8	0	0	0	0	0	5	
A730098P...	2.94	2.15	-1.46	-1.50	3.32	1	5134	0	2	0	2	0	0	0	1	
A930003A...	0.00	0.00	0.00	1.00	0.00	0	0	12	12	0	0	0	0	0	0	
A1406533...	1,421.74	1,418.68	-1,401.65	-41.02	1,455.23	1	2932	552	553	501	501	13	13	2	4	
Abcd5	2,226.79	2,217.07	-2,180.73	-143.13	2,154.88	2	6197	1849	1850	1568	1568	186	186	10	32	
Abcf3	2,696.48	2,690.88	-2,680.59	-4.48	3,437.14	1	3288	1358	1358	1327	1327	270	270	5	21	
Adipoq	13.18	6.27	7.12	3.06	6.91	1	1232	2	2	1	1	0	0	0	3	
Ahsg	348,713.63	345,170.55	336,148.43	11.32	30,789.75	1	1474	5352	5352	5329	5329	473	473	2	7	
Alg3	1,135.93	982.70	-947.13	-2.32	1,550.65	1	1406	267	267	256	256	33	33	2	9	
Ap2m1	40,435.35	39,963.32	-38,612.70	-10.03	44,474.27	1	2057	10768	10769	10741	10742	1142	1142	6	12	
Appd	27,749.25	27,747.89	-27,421.25	-4,638.68	27,999.83	1	1862	9957	9958	9924	9925	555	555	17	6	
Arvcf	1,409.28	1,369.24	-1,344.46	-7.05	1,611.57	1	4624	902	938	844	875	81	83	1	19	
Atp13a3	469.46	400.75	373.99	2.22	273.00	2	7331	255	255	235	235	24	24	0	35	
Atp13a4	327.18	325.97	-307.03	-507.49	288.09	1	4050	197	197	137	137	20	20	3	30	
Atp13a5	229.17	228.04	-198.96	-350.37	169.89	1	4311	108	108	86	86	15	15	0	30	
B3gnt5	31.57	30.16	-25.73	-12.22	22.88	1	4095	11	11	11	11	0	0	0	4	
BC022623	1,190.46	1,155.65	-1,146.30	-10.67	1,291.65	1	3033	460	460	460	460	0	0	0	1	
BC052055	954.15	954.15	-910.99	-∞	954.15	1	2401	329	329	269	269	28	28	12	15	
BC106179	0.00	0.00	0.00	1.00	0.00	0	0	6	6	0	0	0	0	0	0	
Bcf6	1,890.22	1,828.22	-1,800.04	-4.19	2,304.75	1	3285	929	929	889	889	73	73	2	10	
Bdh1	1,072.59	942.30	803.75	1.16	5,118.90	1	1697	1272	1272	1020	1020	97	97	1	7	
Camk2n2	23,041.25	22,995.89	-22,606.68	-296.47	23,095.08	1	1277	3503	3503	3463	3463	92	92	2	2	
Ccdc50	100.41	76.12	-55.79	-1.25	245.03	2	3580	125	125	103	103	10	10	0	11	
Cdc45	177.53	163.64	-154.92	-8.51	191.20	1	2138	55	55	48	48	10	10	0	20	
Cemb2	331.54	302.78	-286.25	-3.56	367.26	1	6493	425	426	280	280	19	19	3	23	
Chrd	1,170.63	1,166.84	-1,102.25	-29.55	1,207.24	1	3273	462	532	413	464	65	66	11	12	
Cln2n	1,875.69	1,759.09	-1,810.67	-4.15	2,377.95	1	3123	997	997	872	872	153	153	19	24	
Cldn1	705.55	684.54	591.06	5.40	123.69	1	3236	48	48	47	47	2	2	0	4	
Cldn16	0.00	0.00	0.00	1.00	0.00	1	1150	2	2	0	0	0	0	0	5	
Cldn5	8,419.05	8,290.93	-8,019.86	-30.07	8,630.87	1	1414	1433	1433	1433	1433	0	0	0	1	
Comt	9,108.91	8,733.74	7,690.02	1.67	11,591.06	1	1252	1736	1789	1651	1704	109	109	2	5	
Cpn2	2,416.88	2,386.87	2,282.27	10.60	252.78	1	2291	71	71	68	68	0	0	0	2	
Crys3	33.30	33.30	-16.65	0.00	0.00	0	0	0	0	0	0	0	0	0	3	
D16H225...	6,984.96	6,968.80	-6,872.52	-6.79	8,163.27	1	1471	1464	1465	1410	1410	168	168	2	8	
Dgcr6	5,248.02	5,110.27	-4,900.37	-4.79	6,471.68	1	1262	1215	1215	959	959	89	89	14	14	
Dgcr8	968.29	952.50	-915.45	-7.39	1,013.67	1	4226	539	540	503	503	40	40	1	14	
Dgkq	2,824.61	2,822.58	-2,756.20	-2,717.47	2,689.82	1	3201	1138	1139	1011	1011	169	169	3	24	
Dnajb11	1,562.27	1,444.53	-1,425.17	-2.24	2,656.94	1	2513	1054	1054	784	784	127	127	27	11	
Dv3	1,688.81	1,660.54	-1,546.69	-8.85	1,871.71	1	2953	669	669	649	649	91	91	3	15	

Sample level

- Expression values
- Transcripts annotated
- Exon length
- Unique gene reads
- Total gene reads
- Unique exon reads
- Total exon reads
- Unique exon-exon reads
- Total exon-exon reads
- Unique intron-exon reads
- Total intron-exon reads
- Exons
- Putative exons
- RPKM
- Median coverage
- Chromosome region start
- Chromosome region end



Expression Analysis: Log 変換



- Navigation Areaから使用するExperimentデータを選択。
- Toolboxから Transform 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。

Expression Analysis: Log 変換

Transform

1. Select either samples or an experiment
2. Set parameters

Set parameters

Values to analyze

- Original expression values
- Transformed expression values
- Normalized expression values

Transformation method

- Logarithm transformation
 - Log 2
- Add a constant
 - 0.0
- Square root transformation

? [Help Icon] Previous Next Finish Cancel

- Value to analyze: 解析に使用したい値
 - Original expression values
 - Transformed expression values
 - Normalized expression values
- Transformation method
 - Log2, Log10, Log e
 - 任意の数値を使った変換
 - 平方根



Expression Analysis: Log 変換

Brain vs. Liver x

Rows: 181 Filter: Match any Match all

Feature ID contains

Feature ID	Experiment									Express...	Transformed values	Transcrip...	Exon length	Unique ge...	Total gen...	Unique ex...	Total exo...	Uni
	Range (or...	IQR (origi...	Differenc...	Fold Chan...	Range (tr...	IQR (tran...	Differenc...	Fold Chan...										
1110054...	599.49	500.74	-509.47	-2.02	1.22	0.93	-1.02	-1.11	1,050.96	10.04	1	705	87	87	87	87	87	
1600021P...	1,422.81	1,421.10	-1,391.08	-6.52	2.74	2.73	-2.71	-1.34	1,612.04	10.65	1	3080	619	620	583	583	583	
1700007E...	0.00	0.00	0.00	1.00	NaN	0.00	NaN	NaN	0.00	-∞	0	0	0	0	0	0	0	
1700025H...	5.94	5.94	2.97	∞	∞	0.00	NaN	NaN	0.00	-∞	1	547	0	0	0	0	0	
1810009K...	337.32	335.49	313.41	2.17	1.18	1.17	1.12	1.14	268.30	8.07	1	984	35	35	31	31	31	
2310042E...	18.35	16.42	-14.89	-16.47	∞	3.25	-∞	0.00	13.35	3.74	1	1276	0	2	0	2	2	
2510002D...	494.02	466.47	-455.21	-2.16	1.21	1.11	-1.11	-1.13	871.00	9.77	1	1232	160	160	126	126	126	
2510009E...	1,402.24	1,401.64	-1,377.14	-22.15	4.50	4.49	-4.47	-1.74	1,417.46	10.47	1	5113	885	885	851	851	851	
2900046G...	15,812.44	15,806.41	-15,335.32	-1,821.61	11.51	10.43	-10.93	-4.67	15,817.85	13.95	1	2271	4275	4277	4216	4218	4218	
5430420C...	4.32	4.32	-2.16	-∞	∞	0.00	NaN	NaN	0.00	-∞	1	1084	28	28	0	0	0	
6430590I...	0.00	0.00	0.00	1.00	NaN	0.00	NaN	NaN	0.00	-∞	0	0	52	52	0	0	0	
9030404E...	0.00	0.00	0.00	1.00	NaN	0.00	NaN	NaN	0.00	-∞	1	488	8	8	0	0	0	
A730098F...	2.94	2.15	-1.46	-1.50	1.11	0.72	-0.55	-1.36	3.32	1.73	1	5134	0	2	0	2	2	
A930003A...	0.00	0.00	0.00	1.00	NaN	0.00	NaN	NaN	0.00	-∞	0	0	12	12	0	0	0	
A1480653	1,421.74	1,418.68	-1,401.65	-41.02	5.44	5.32	-5.36	-2.04	1,453.23	10.51	1	2932	552	553	501	501	501	
Abcc5	2,226.79	2,217.07	-2,180.73	-143.13	7.74	6.79	-7.24	-2.87	2,154.88	11.07	2	6197	1849	1850	1568	1568	1568	
Abcf3	2,696.48	2,690.88	-2,680.59	-4.48	2.18	2.16	-2.16	-1.23	3,437.14	11.75	1	3288	1358	1358	1327	1327	1327	
Adipoq	13.18	6.27	7.12	3.06	∞	0.93	∞	-0.00	6.91	2.79	1	1232	2	2	1	1	1	
Ahsg	348,713.63	345,170.55	336,148.43	11.32	3.62	3.47	3.50	1.23	30,789.75	14.91	1	1474	5352	5352	5329	5329	5329	
Alec3	1,135.93	982.70	-947.13	-2.32	1.47	1.16	-1.22	-1.13	1,550.65	10.60	1	1406	267	267	256	256	256	

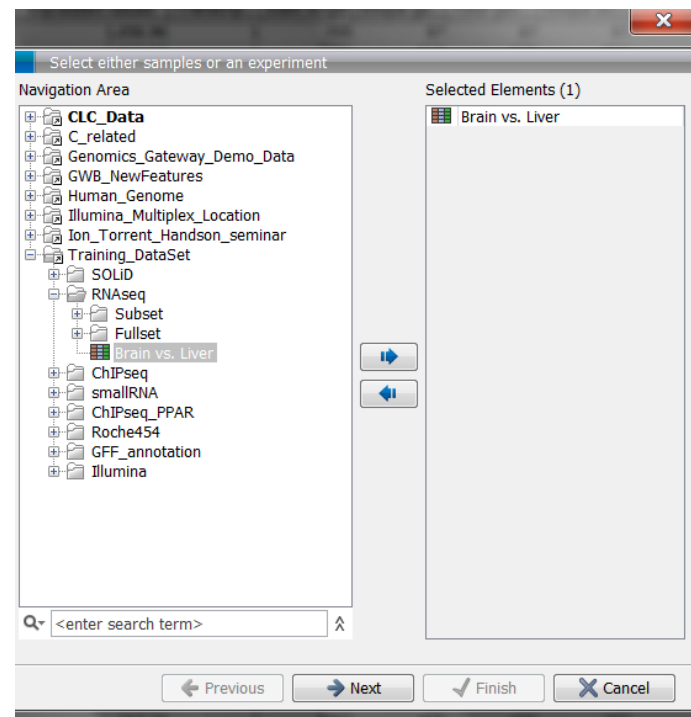
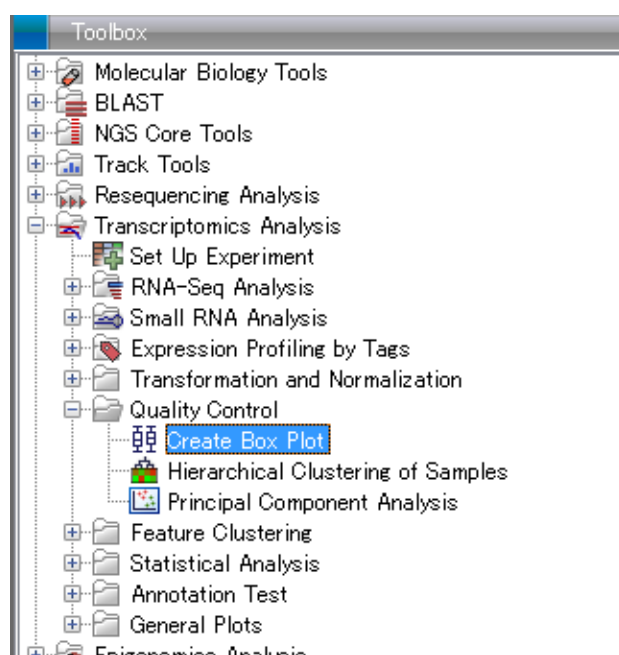
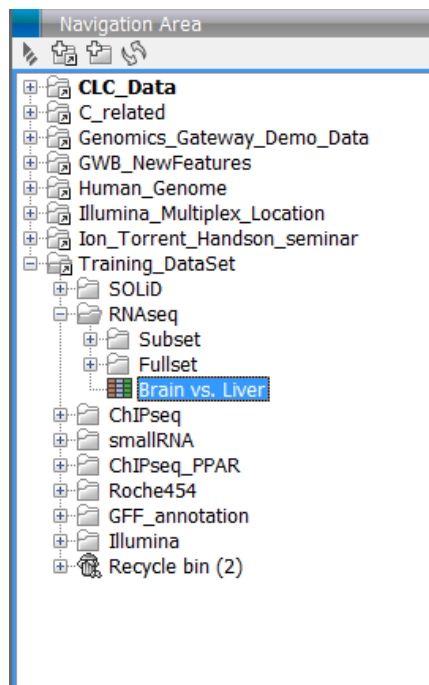
- Sample level
 - Expression values
 - Transformed values
 - Transcripts annotated
 - Exon length
 - Unique gene reads
 - Total gene reads

■ 変換された値が表に追加される



Expression Analysis: Box Plot

- Log変換後の結果をBox plotで確認

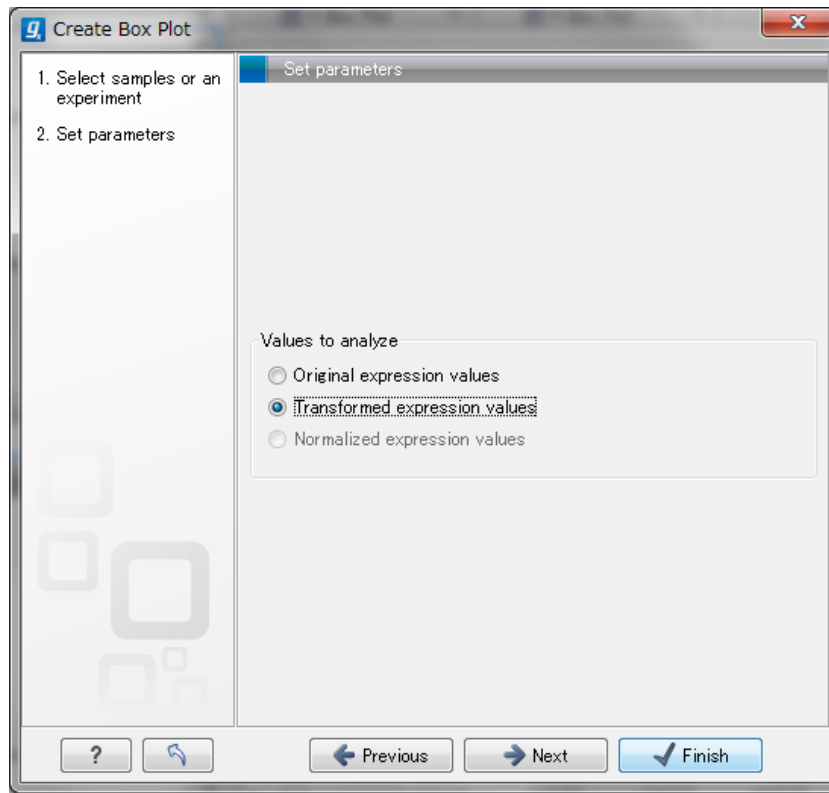


- Navigation Areaから使用するExperimentデータを選択。
- Toolboxから Transcriptomics Analysis > Quality Control > Create Box Plot を選択、ダブルクリック。
- ウィザードが起動し、選択したデータが選ばれていることを確認。



Expression Analysis: Box Plot

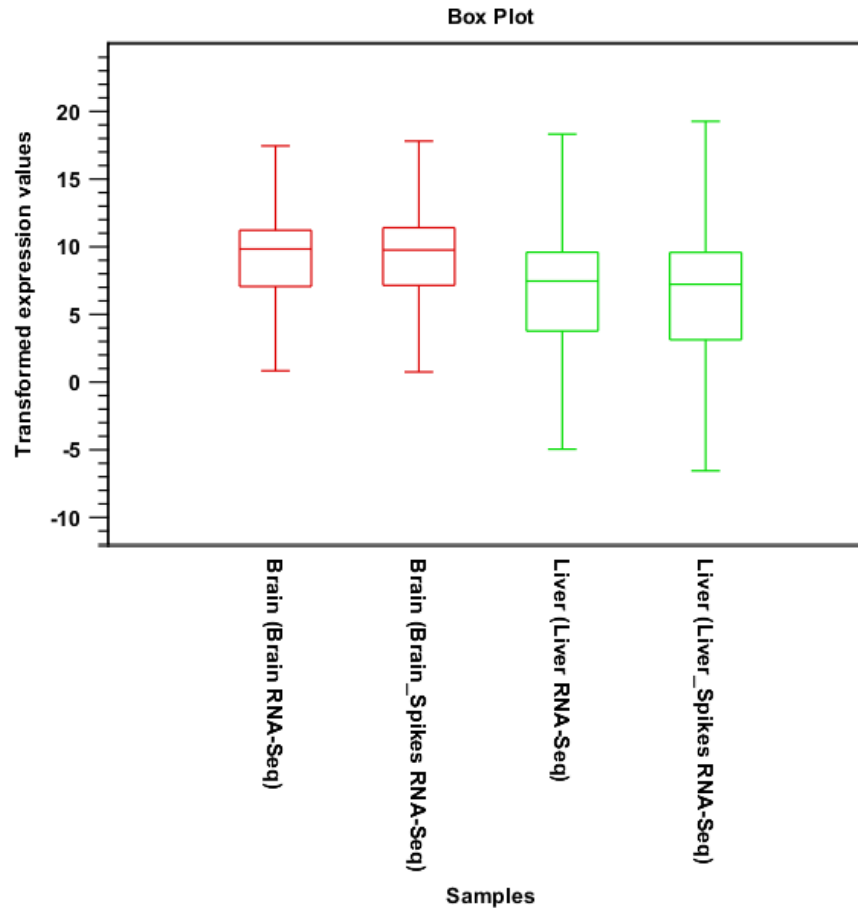
- Log変換後の結果をBox plotで確認



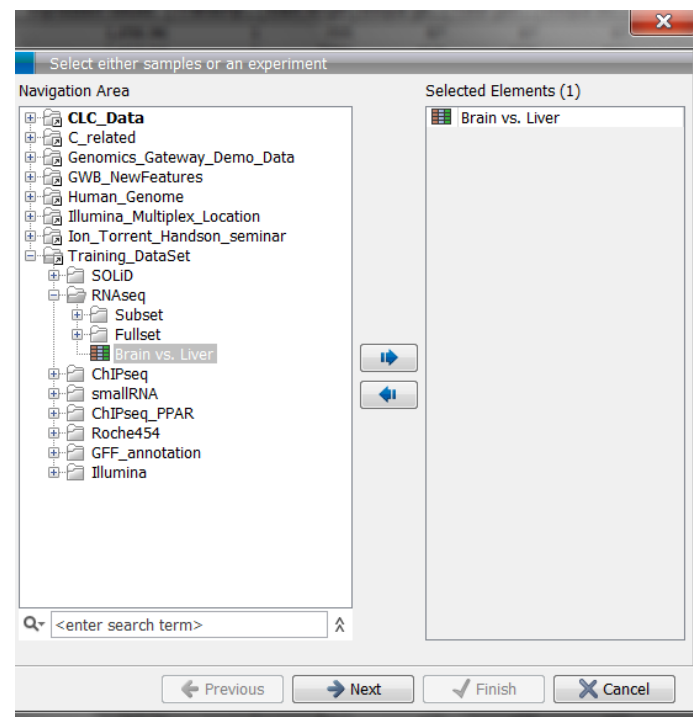
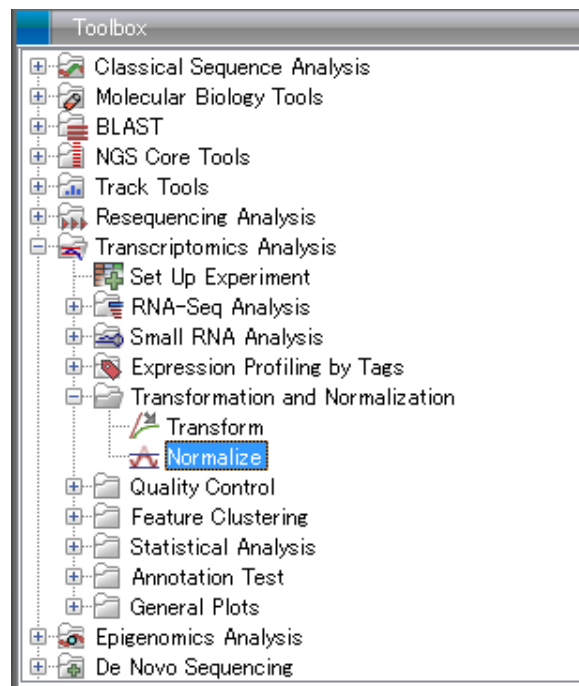
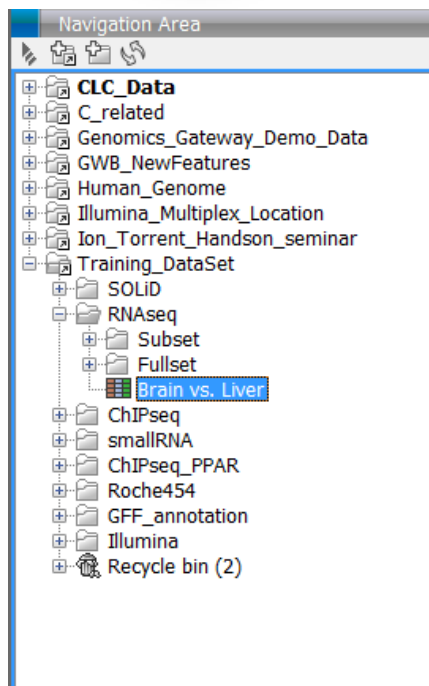
- Value to analyze: 解析に使用したい値を選択
 - もとの発現値
 - 変換後の発現値
 - ノーマライズ後の発現値



Expression Analysis: Box Plot



Expression Analysis: ノーマライゼーション



- Navigation Areaから使用するExperimentデータを選択。
- Toolboxから Transform 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。

Expression Analysis: ノーマライゼーション

1. Select either samples or an experiment

2. Choose normalization method

Choose normalization method

Scaling

Quantile

By totals

state numbers in 'Reads per...':

Values to analyze

Original expression values

Transformed expression values

Normalized expression values

? Refresh Previous Next Finish Cancel

- Choose normalization method
 - Scaling: ある固定の値でノーマライズ。この選択肢を選ぶと、ノーマライズする値について次のウィンドウで選択する。
 - Quantile: グループ間で分布が同じ形から来ていると仮定し、ノーマライズする。
 - By total: 発現値をカウントの値としたときに使用する。
- Value to analyze: 解析に使用したい値を選択
 - もとの発現値
 - 変換後の発現値
 - ノーマライズ後の発現値

Expression Analysis: ノーマライゼーション

- Scaling を選んだ場合の次の画面

1. Select either samples or an experiment

2. Choose normalization method

3. Set parameters

Set parameters

Choose normalization value

Mean

Median

Choose reference

Median mean

Median median

Use another sample

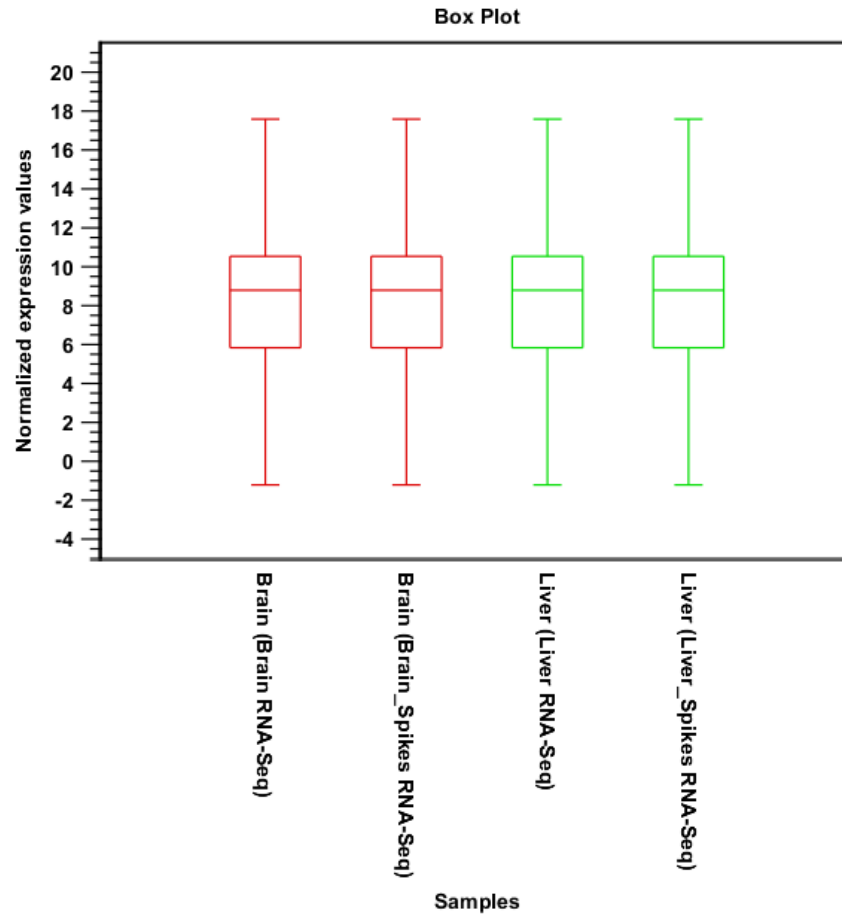
Trimming

Trimming percentage

?

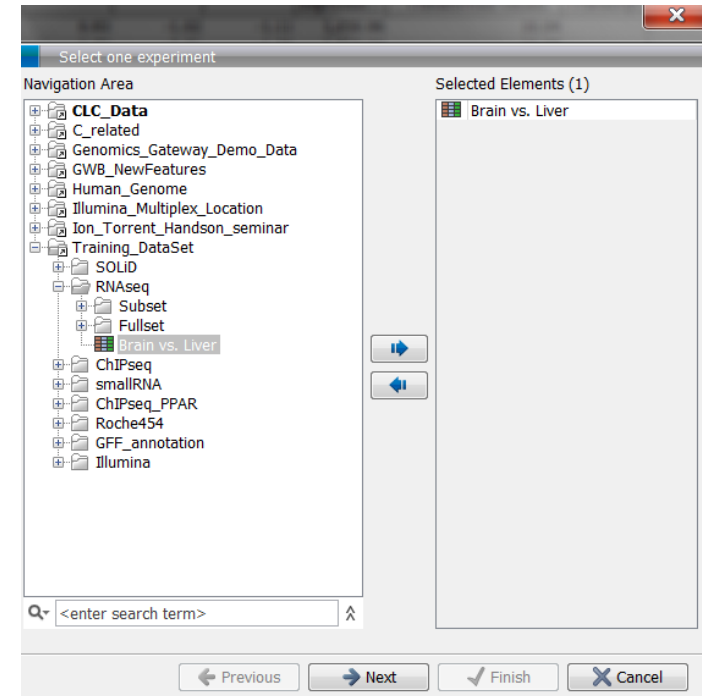
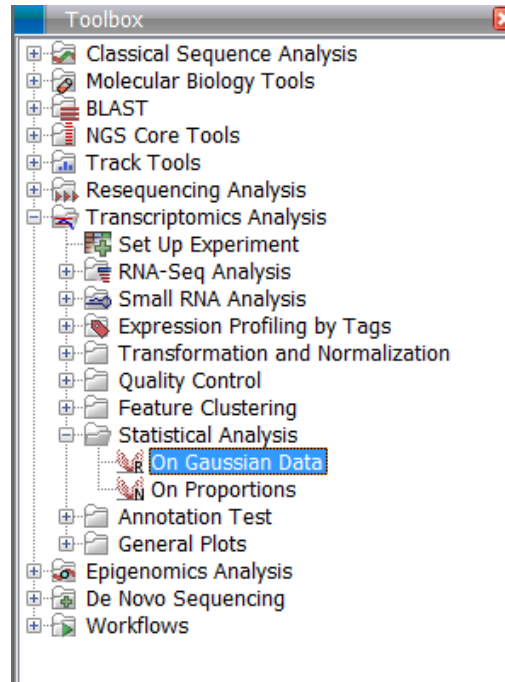
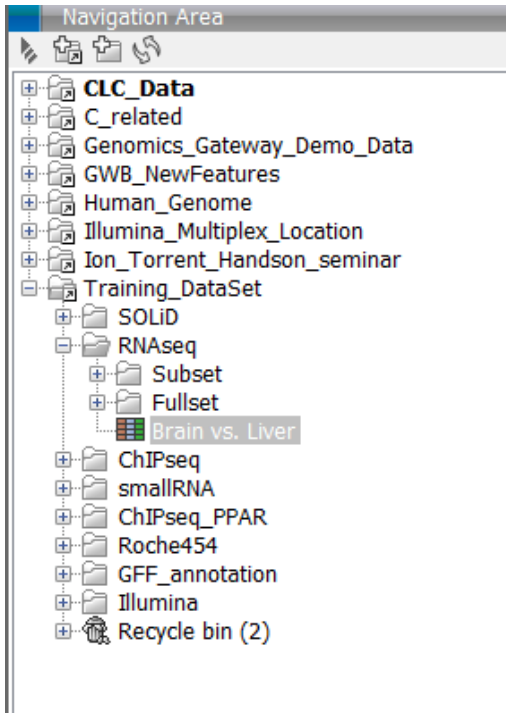
- Choose normalization value: ノーマライズ後にそろえる値を平均値か中央値を選択
- Choose reference: ノーマライズに使用する値。トリミング後の値を使用するが、トリミング後の値の平均値をつかうか中央値を使うか決める。
- Trimming: トリムする%を入力

Expression Analysis: ノーマライゼーション





Expression Analysis: t-test



- Navigation Areaから使用するExperimentデータを選択。
- Toolboxから On Gaussian Data 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。



Expression Analysis: t-test

On Gaussian Data

1. Select one experiment
2. Statistical analysis

Statistical analysis

Choose statistic

T-test

Variances:

Homogeneous

In-homogeneous

Comparisons:

All pairs

Against reference Brain

ANOVA

Pairing

Use pairing

? Refresh Previous Next Finish Cancel

On Gaussian Data

1. Select one experiment
2. Statistical analysis
3. Set parameters

Set parameters

Values to analyze

Original expression values

Transformed expression values

Normalized expression values

Add corrected p-values

Bonferroni corrected

FDR corrected

? Refresh Previous Next Finish Cancel

■ t検定

■ 分散

■ 均一な場合

■ 不均一

■ 全てのペアで比較するか、任意のグループに対して比較するか。

■ Value to analyze

■ もとの発現値

■ 変換後の発現値

■ ノーマライズ後の発現値

■ p値の補正

■ ボンフェローニ

■ FDR



Expression Analysis: t-test

Rows: 181 Filter:

			t-test: Brain vs Liver normalized values					
IQR (normal...	Difference (...	Fold Chang...	Difference	Fold change	Test statistic	P-value	FDR p-valu...	Ex
0.00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
0.79	0.83	1.08	0.83	1.08	4.33	0.08	0.25	
0.00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
0.74	0.81	1.08	0.81	1.08	5.07	0.11	0.30	
0.10	-0.10	-1.01	-0.10	-1.01	-1.41	0.29	0.69	
0.67	-0.63	-1.06	-0.63	-1.06	-4.38	0.09	0.28	
1.90	1.88	1.18	1.88	1.18	10.32	0.01	0.17	
0.32	-0.32	-1.03	-0.32	-1.03	-∞	NaN	NaN	
0.37	-0.39	-1.04	-0.39	-1.04	-17.06	0.04	0.18	
2.27	2.48	1.45	2.48	1.45	11.91	0.05	0.20	
8.33	-∞	0.00	-∞	0.00	NaN	NaN	NaN	
1.62	-1.58	-1.78	-1.58	-1.78	-6.21	0.04	0.18	
0.00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
4.16	4.16	1.77	4.16	1.77	59.49	6.20E-4	0.11	
0.07	-0.16	-1.02	-0.16	-1.02	-1.82	0.32	0.74	
10.66	-10.69	-2.64	-10.69	-2.64	-318.49	2.00E-3	0.12	
0.94	0.90	1.09	0.90	1.09	8.51	0.05	0.20	



その他の機能について

- その他の機能については、弊社ホームページから日本語資料をダウンロードいただけます。

<http://www.clcbio.co.jp/index.php?id=223>

Genomics Workbench サポート資料















更新日2012年8月10日

トレーニング用資料

ここでは製品の利用に関する資料を掲載しています。ご利用に際し、パラメータの詳細な説明などを日本語で記載しています。さらに詳細なマニュアル(英語)については、こちらをご参照ください。

Version 6.0

ダウンロード

	データインポート	
	クオリティチェック	
	マッピング	
	SNV検出	
	SNV比較とアノテーション	
	RNA-seq	
	ChIP-seq	

ご清聴ありがとうございました。



APPENDIX



P値の補正

- 検定を多く繰り返す(たくさんの遺伝子を一度に検定する)と多くのエラーを含んだリストを返す結果となります。
- たとえば、 $p < 0.05$ 以下の遺伝子のリストをえたい場合、3つの遺伝子をそれぞれケースとコントロールで検定した結果のリストは $1 - (1 - 0.05)^3 = 0.14$ となり、実際に得られるリストは、p値が0.05以下のリストではなく、0.14以下のリストとなります。
- ボンフェローにはこれを抑えるため、設定するp値を検定する数(発現解析では遺伝子の数、上記の例では3)で割り、小さなp値の閾値でリストを取得します。



P-value correction

FDR

- Say $p_1 < p_2 < p_3 < \dots < p_i < \dots < p_m$ and α is threshold.
- $i = m$
- If $p_i < \alpha \frac{i}{m}$ (1) を満たすならば、 $k = i$
- (1)式が満たされない場合 $i = m - 1$ として (1)を再度計算
- p_1, \dots, p_k に対応する仮説を棄却する。



カウントデータの検定

Kal's test

- 2つのグループのカウントデータを比較し、その差が統計的に有意かどうかを検定する手法。

$$Z = \frac{p_A - p_B}{\sqrt{\frac{p_0(1-p_0)}{N_A} + \frac{p_0(1-p_0)}{N_B}}}, \quad \text{with}$$

Aの分散 $\frac{p_0(1-p_0)}{N_A}$ Bの分散 $\frac{p_0(1-p_0)}{N_B}$

$$p_A = \frac{X_A}{N_A}, \quad p_B = \frac{X_B}{N_B}, \quad p_0 = \frac{X_A + X_B}{N_A + N_B},$$

where the two library sizes are N_A and N_B and the two counts are X_A and X_B , respectively.

Kal, A. J. *et al.* Dynamics of Gene Expression Revealed by Comparison of Serial Analysis of Gene Expression Different Carbon Sources. **10**, 1859–1872 (1999).

Baggerly, K. a., Deng, L., Morris, J. S. & Aldaz, C. M. Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics* 19, 1477–1483 (2003).



カウントデータの検定

Baggerley's test

- Kal's test はレプリケートを必要としませんが、レプリケートがあった場合でも、レプリケート内のばらつきを考慮できません。これに対応するため、Baggerleyのテストでは、レプリケート内のばらつきを考慮するために提案された手法です。
- 統計量の算出方法は、Kal'sテストと似ていますが、分散の推定が複雑になっています。

$$t_w = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{V}_A + \hat{V}_B}}$$