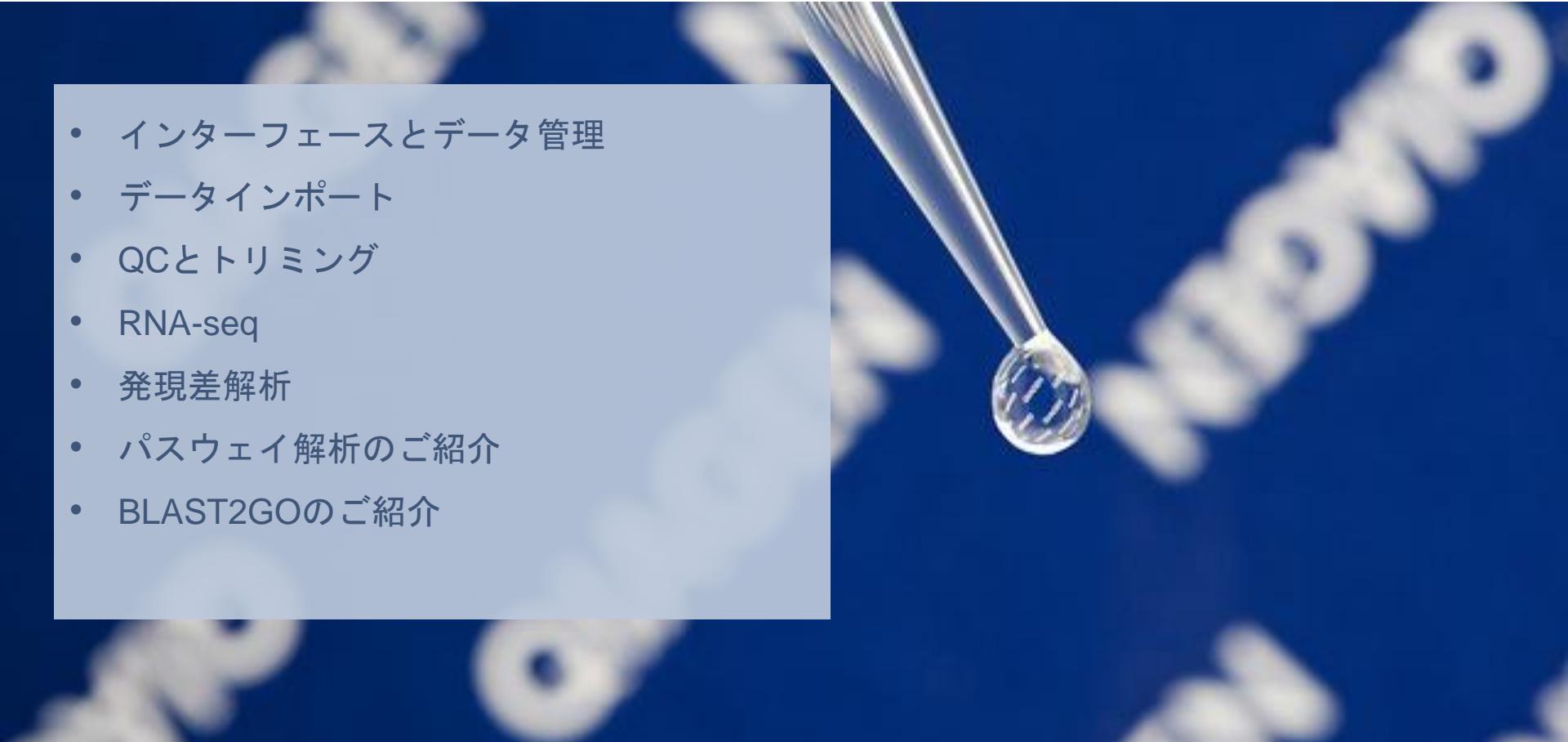




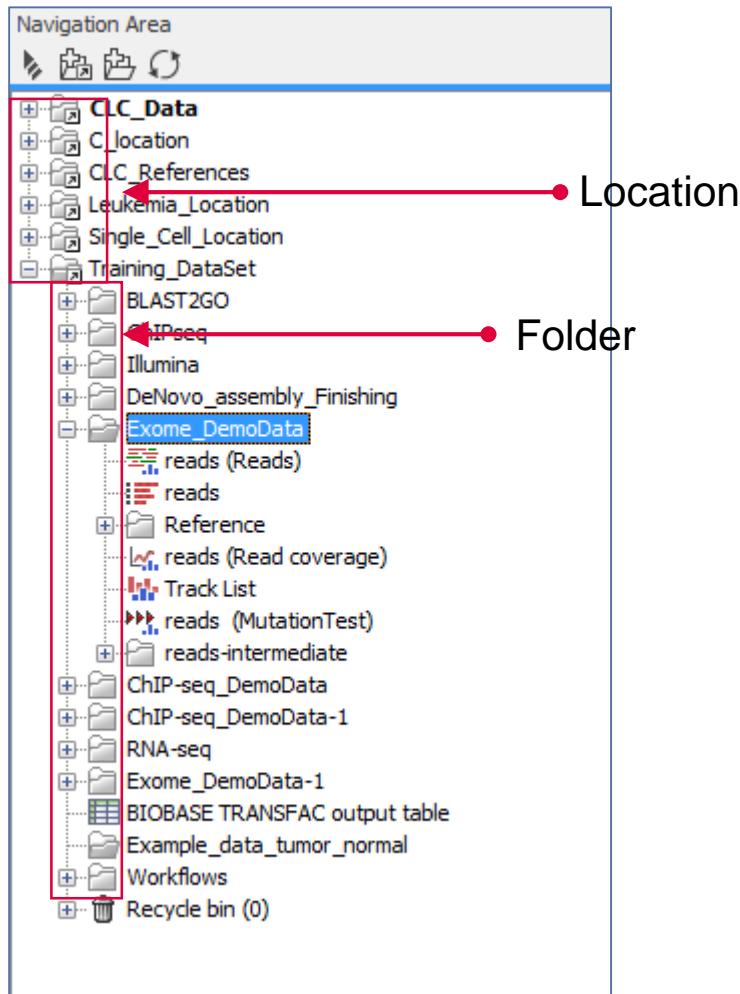
CLC Genomics Workbench ハンズオントレーニング RNA-seq 編

株式会社キアゲン
グローバルイフオマティクス 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
宮本 真理
mari.miymoto@qiagen.com

- インターフェースとデータ管理
- データインポート
- QCとトリミング
- RNA-seq
- 発現差解析
- パスウェイ解析のご紹介
- BLAST2GOのご紹介

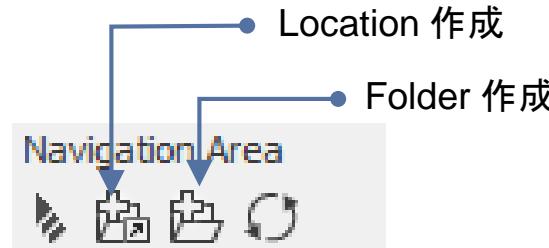


データロケーション



- Genomics Workbench ではデータ保存の階層のトップを Location と呼びます。
- デフォルトの Location は CLC_Data が作成されていますが、左の図のように Location は追加可能です。

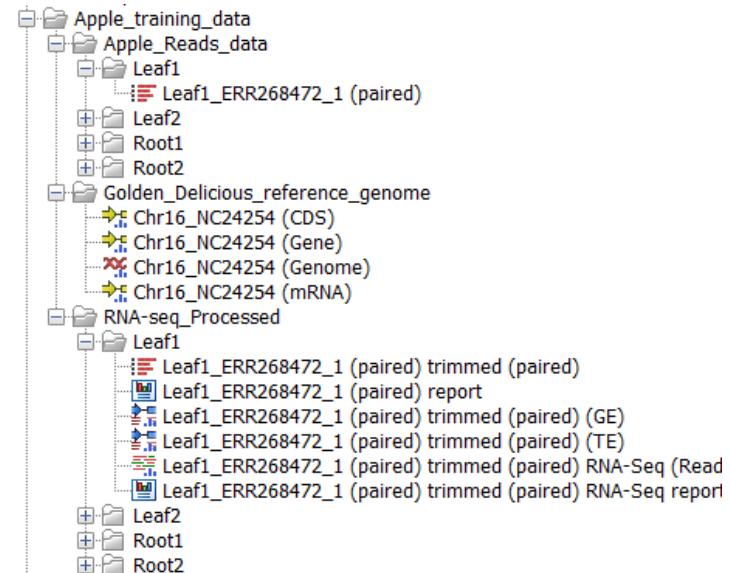
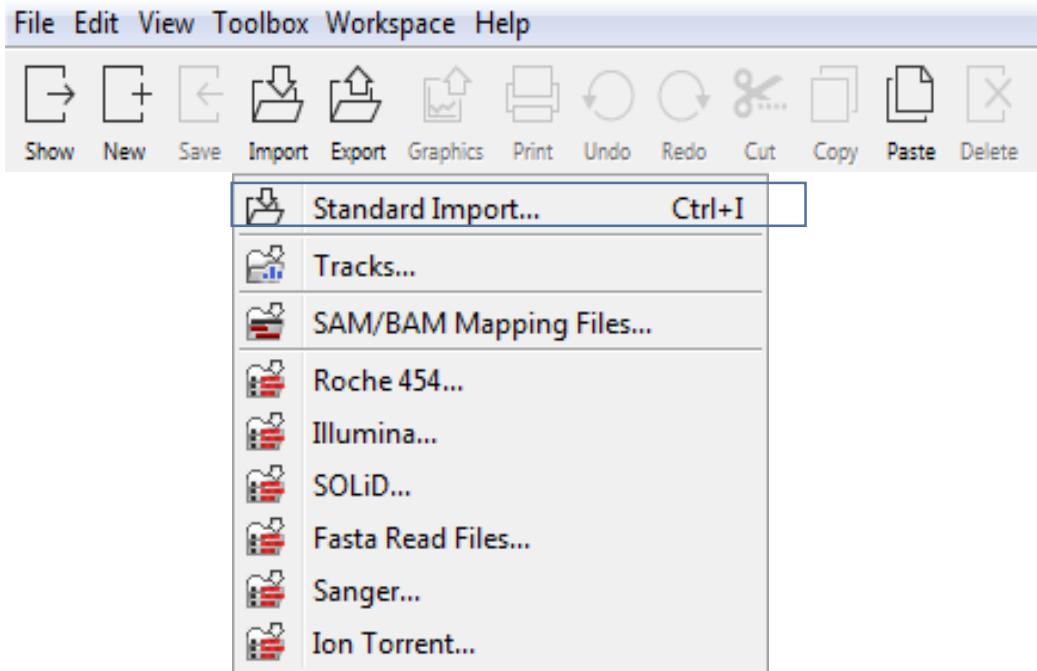
- Location の新規追加は、Navigation Area 左上のアイコンから作成可能です。シークエンスデータはサイズが大きいため、容量が大きいディスクへ Location を作成することをお勧めします。



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

データインポート

今日は、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。

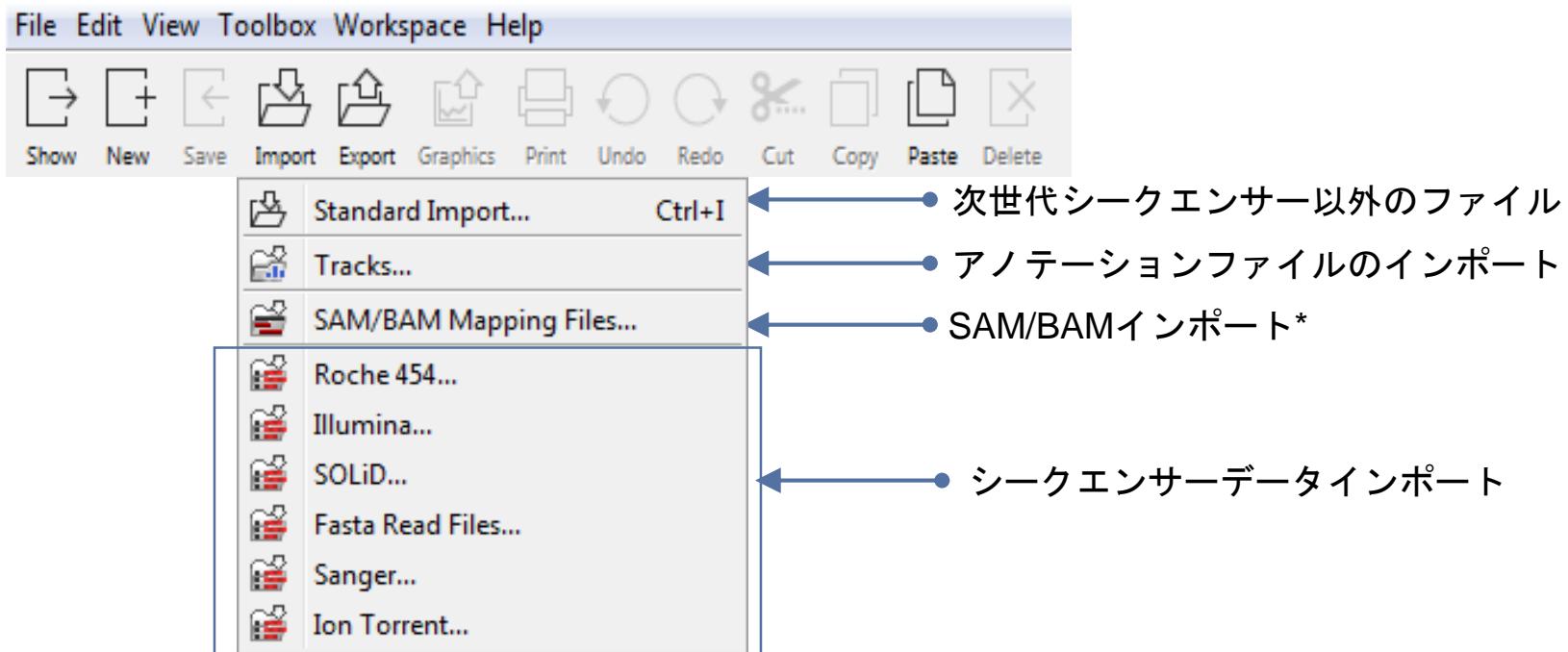


今日のデータはColumnar Apple Tree の根と葉の発現差について、ゴールデンデリシャスの参照ゲノムを使ってみていきます。

CLC Genomics Workbench

データインポート

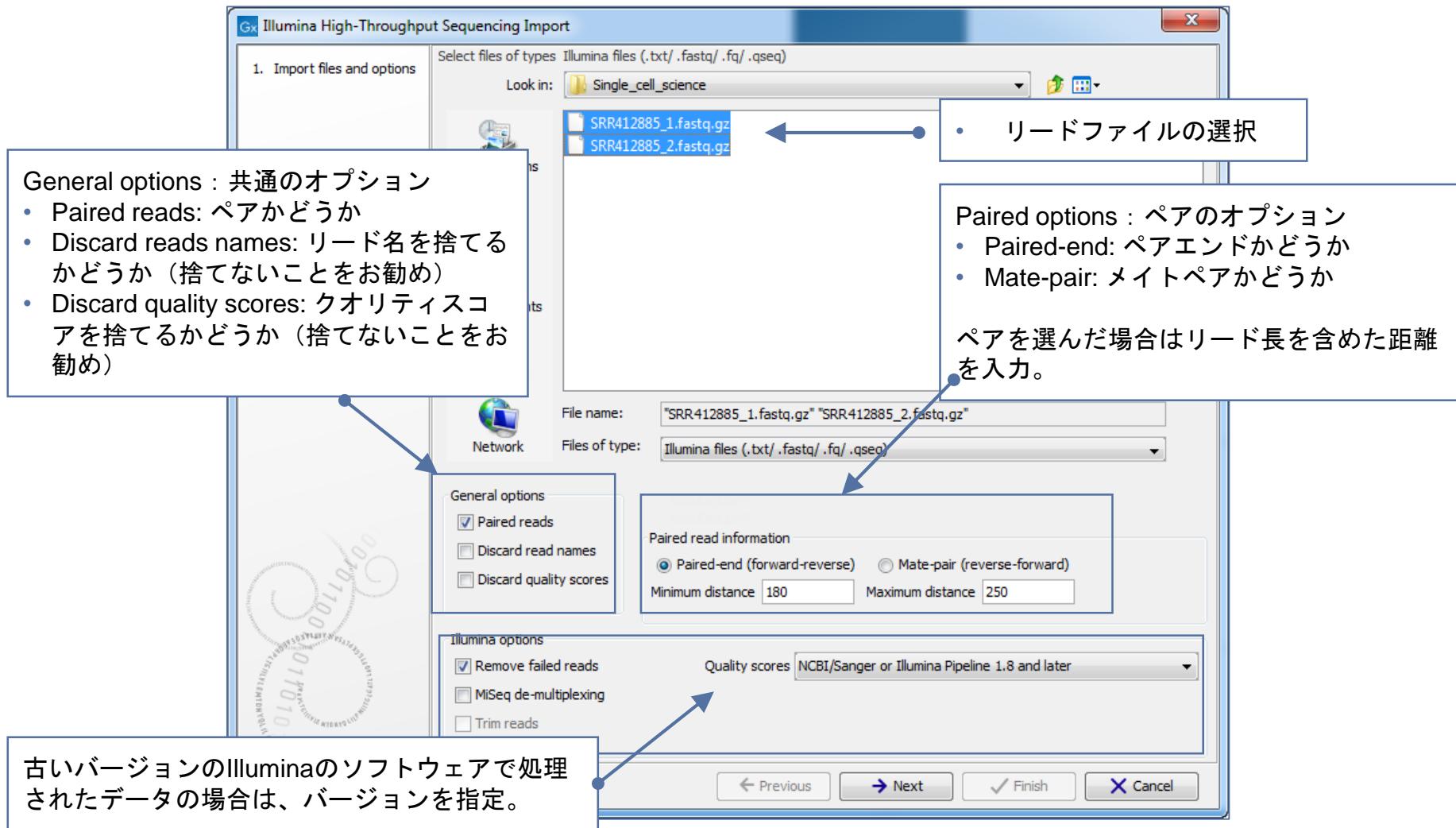
リードデータインポート



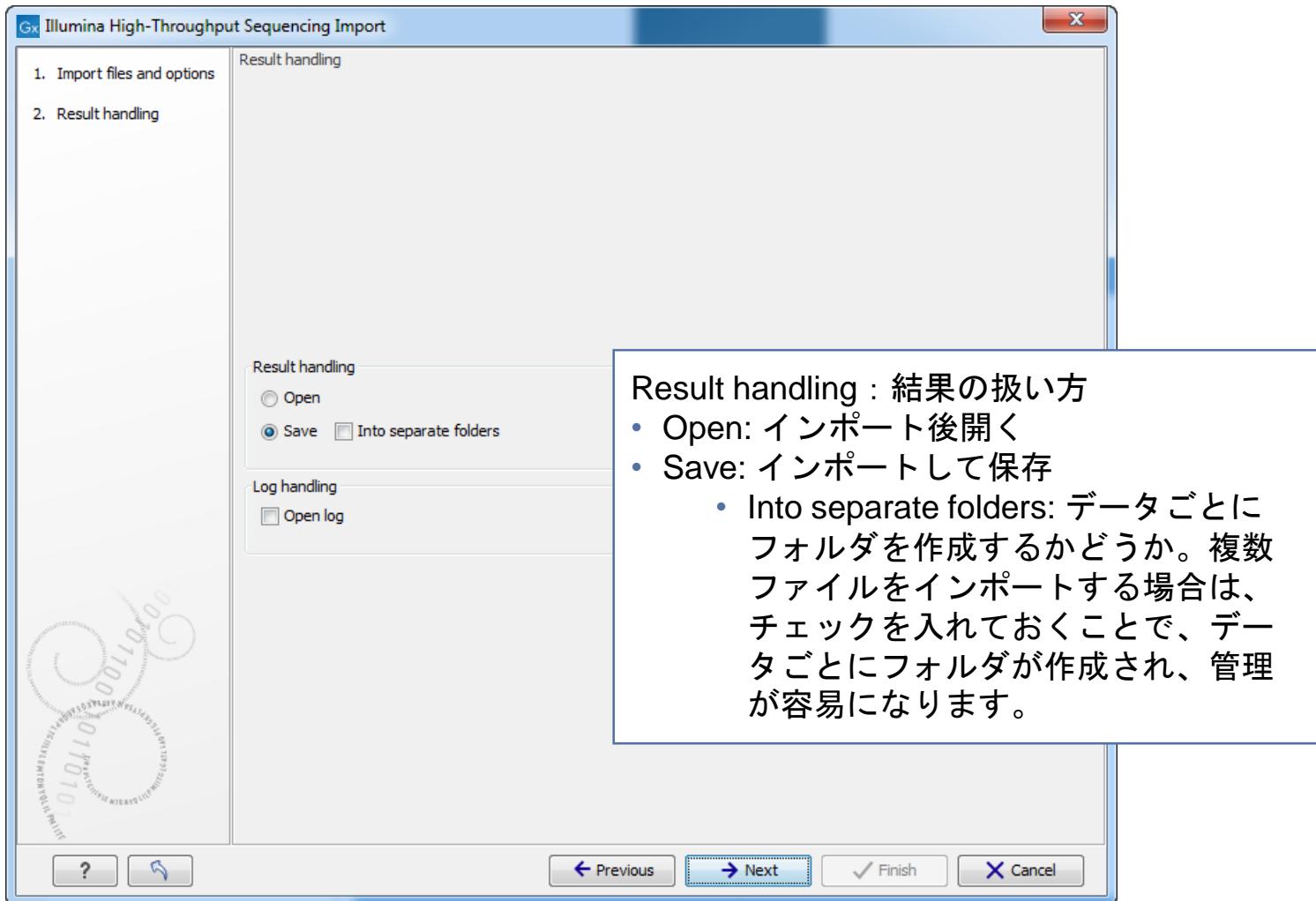
SAM/BAMファイルは、マッピング後のデータにおいて利用される一般的なフォーマットです。

データインポート

リードデータインポート：イルミナ



リードデータインポート：イルミナ



データインポート

リードデータインポート : Ion Torrent

General options : 共通のオプション

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

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

Import files and options

Look in: IonTorrent
File name: STO_408.fastq.gz
Files of type: Ion Torrent (.fastq/ .fq)

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

Minimum distance 100 Maximum distance 500

Ion Torrent options

Use clipping information

Sample to Insight

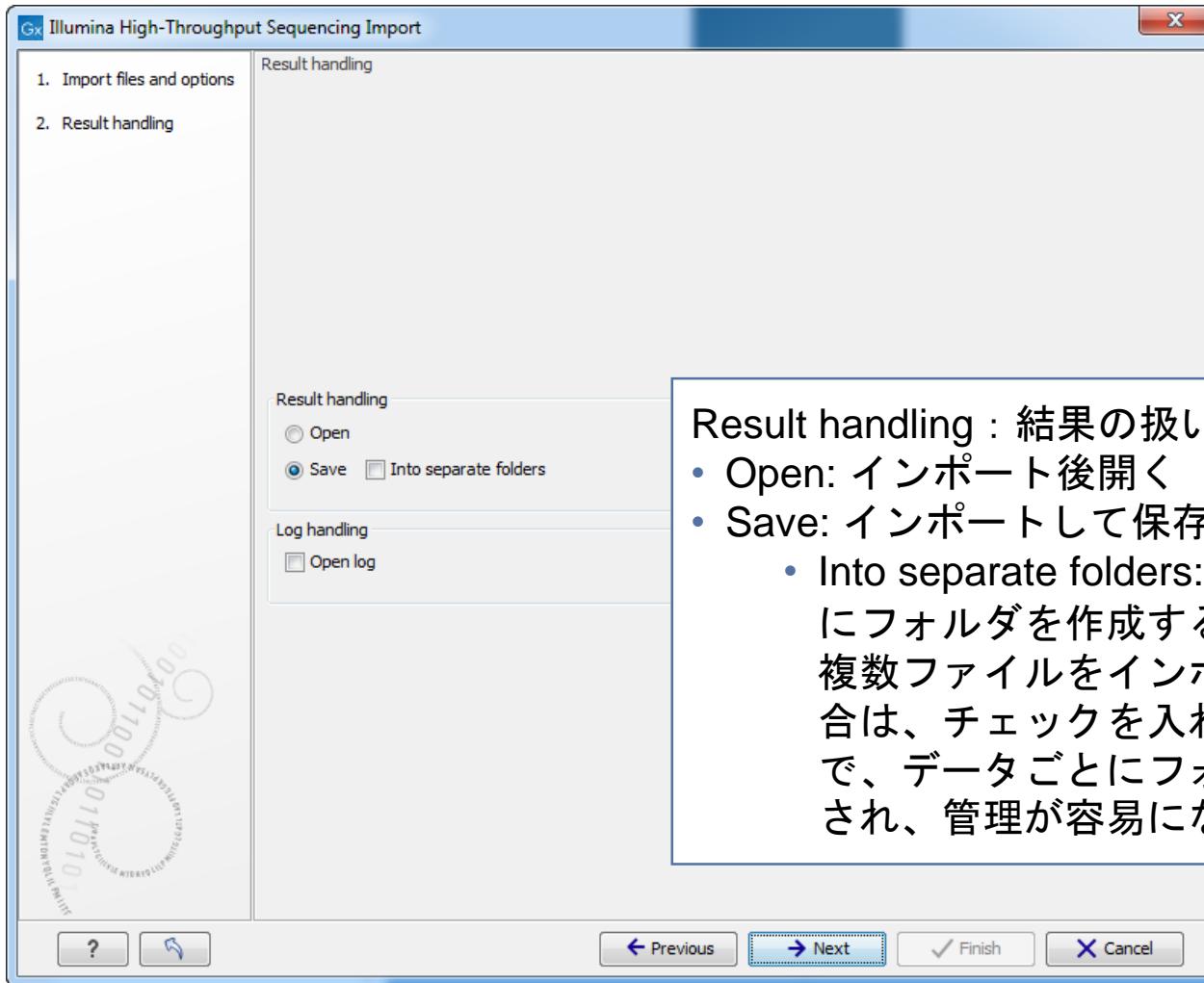
• リードファイルの選択

• Fastqかsffを選択可能

• Paired options : ペアのオプション

ペアを選んだ場合はリード長を含めた距離を入力。

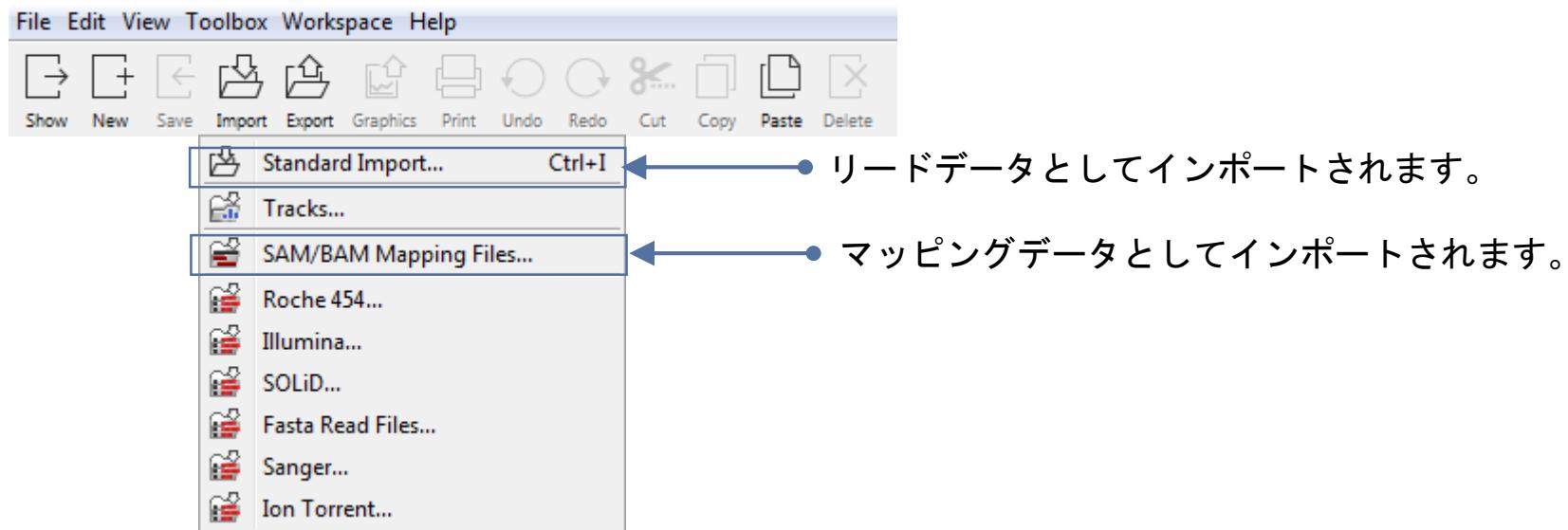
リードデータインポート : Ion Torrent

**Result handling : 結果の扱い方**

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

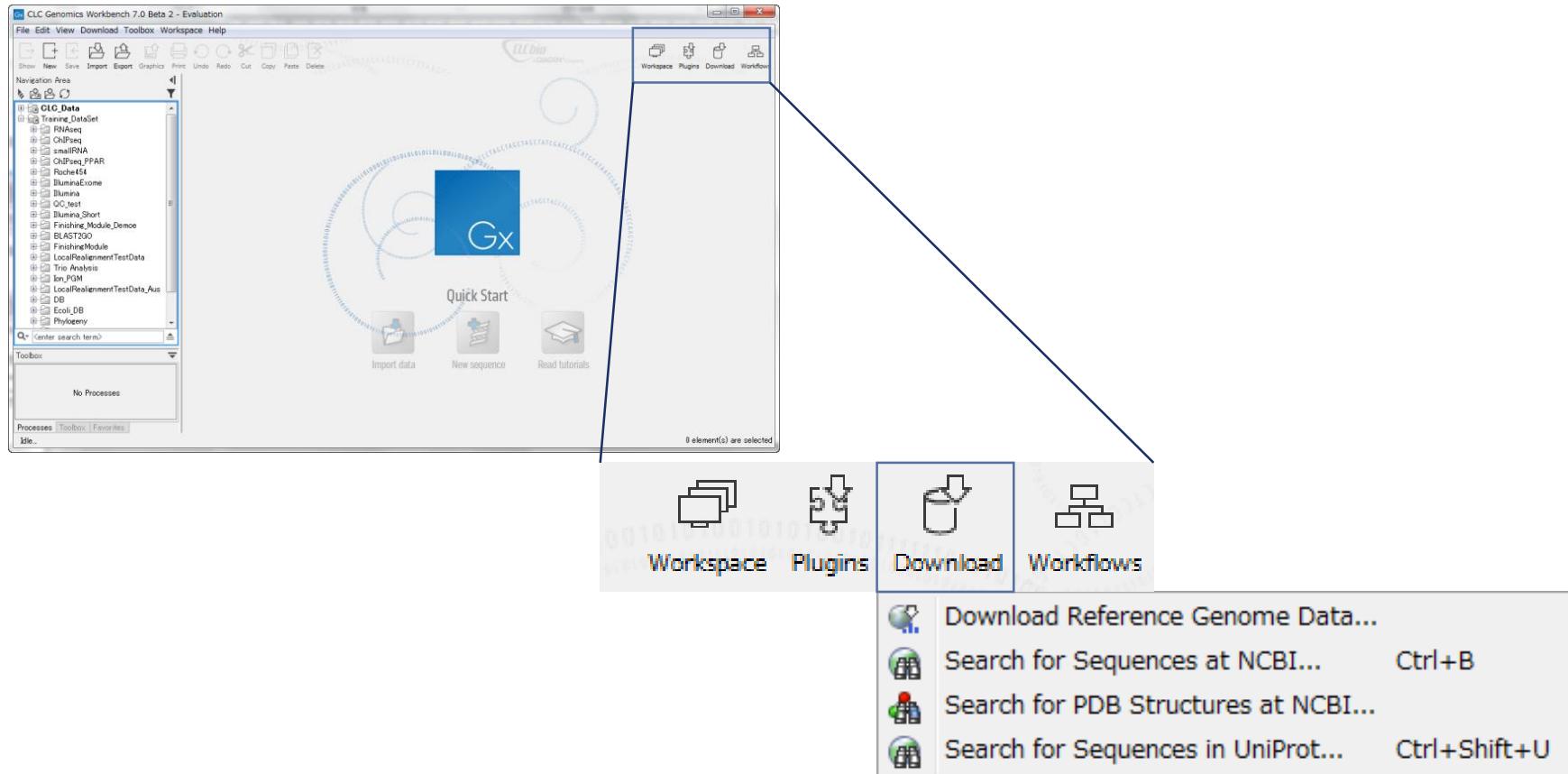
リードデータインポート : Ion Torrent (Unmapped BAMファイル) **※注意**

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

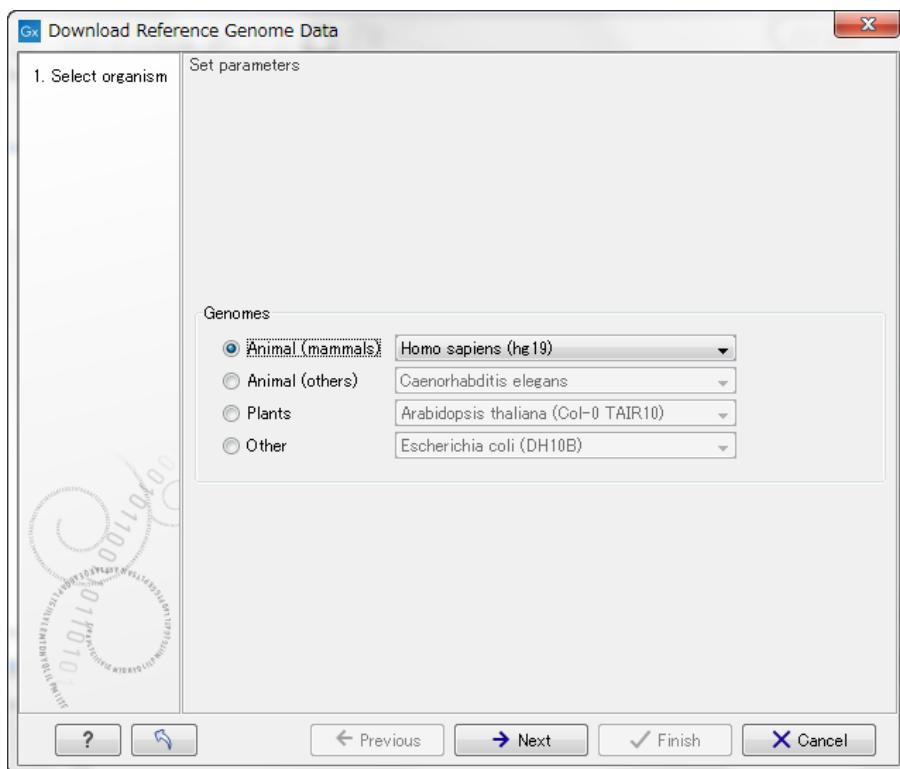


ゲノムインポート

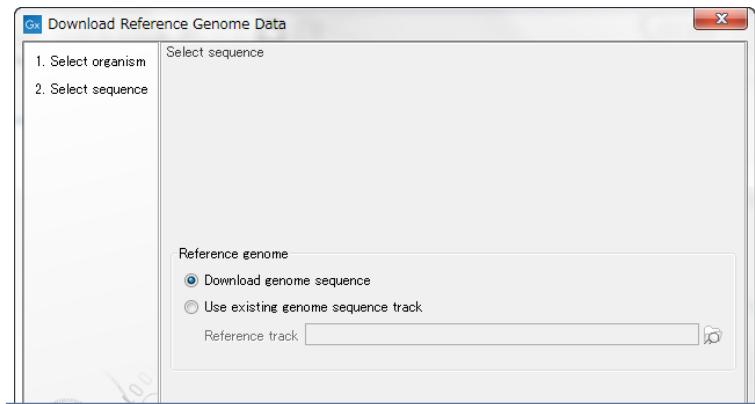
ゲノムデータは、よく知られているモデル動物についてはのDownload Genome よりインポートできます。



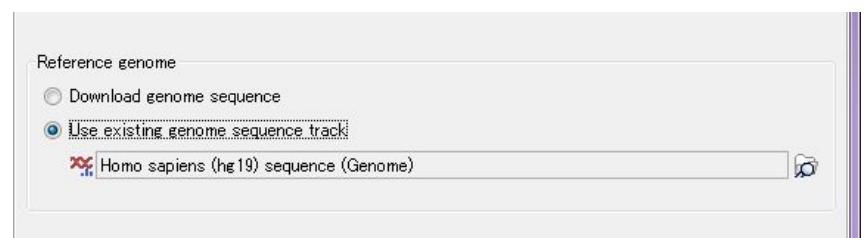
ゲノムインポート



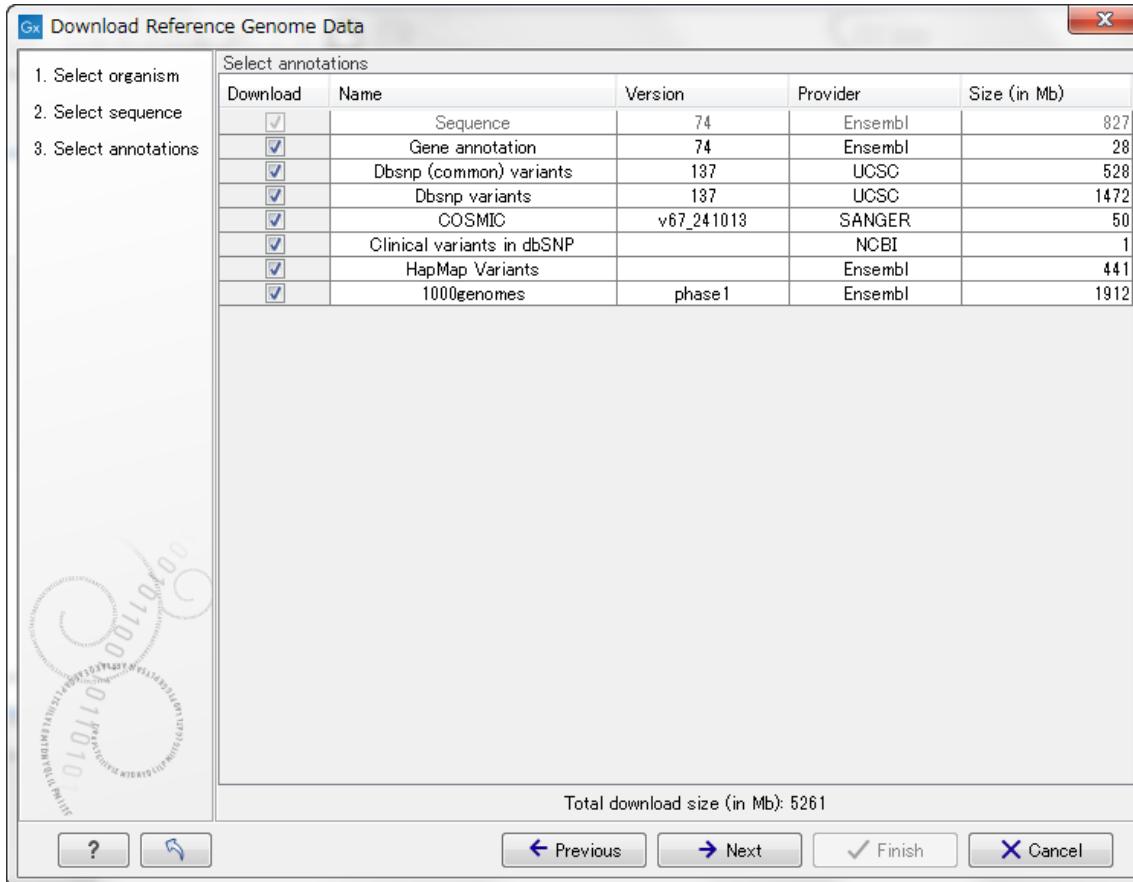
- ドロップダウンリストから生物種を選択。



- 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。

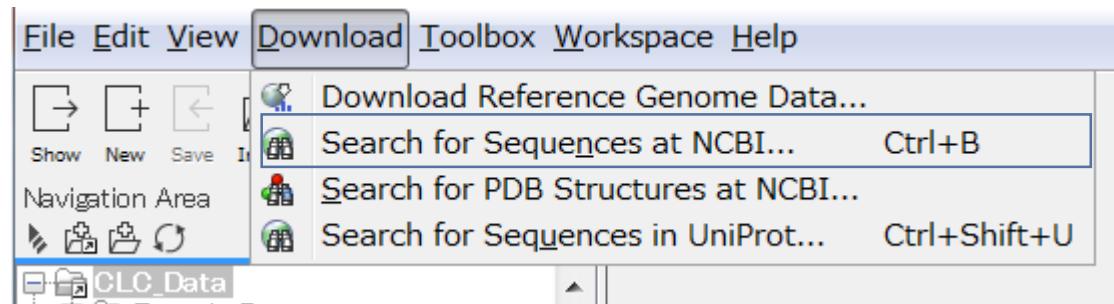


ゲノムインポート

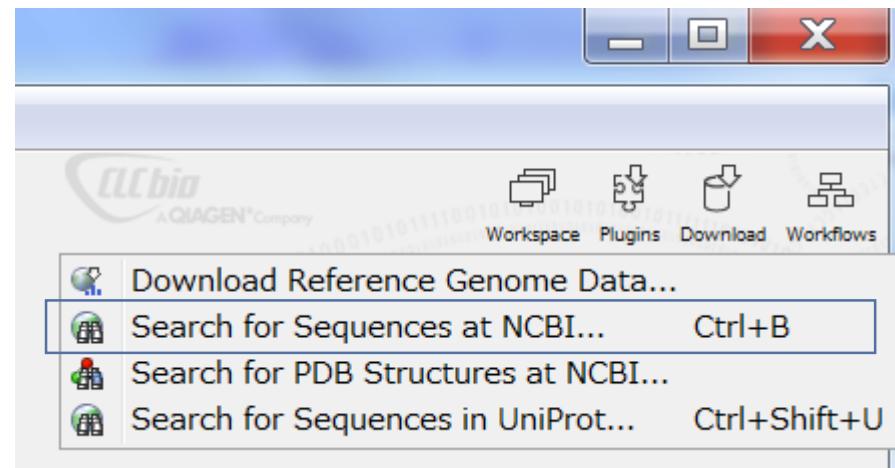


- 希望するアノテーションにチェックを入れる。ゲノム配列をダウンロードするときは、Sequences にもチェックを入れる。
- 選択した生物種により、表示されるアノテーションの種類は異なります。

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

データインポート

Search for Sequences at NCBI

NCBI search

Choose database: Nucleotide Protein

All Fields: Homo sapiens
All Fields: haptoglobin

Add search parameters Append wildcard (*) to search words Start search

Hit	Accession	Description	Modification Date	Length
1	NM_00518	Homo sapiens haptoglobin, beta (HBB), mRNA	2014/02/15	626
2	NM_208416	Homo sapiens CD163 molecule (CD163), transcript variant 2, mRNA	2014/02/08	4107
3	NM_004244	Homo sapiens CD163 molecule (CD163), transcript variant 1, mRNA	2014/02/08	4190
4	XM_006721186	PREDICTED: Homo sapiens haptoglobin-related protein (HPR), transcript variant X1, mRNA	2014/02/08	1119
5	XM_006255922	PREDICTED: Homo sapiens haptoglobin (HP), transcript variant X1, mRNA	2014/02/08	1296
6	NT_01498	Homo sapiens chromosome 16 genomic scaffold, GRCh38 Primary Assembly HSCHR16_0TG3.1	2014/02/08	43847663
7	NC_018827	Homo sapiens chromosome 16, alternate assembly CHM1_1.1, whole genome shotgun sequence	2014/02/08	91765909
8	NW_004929402	Homo sapiens chromosome 16 genomic scaffold, alternate assembly CHM1_1.1, whole genome shotgun sequence	2014/02/08	42007949
9	AC_000148	Homo sapiens chromosome 16, alternate assembly HuRef, whole genome shotgun sequence	2014/02/08	75877710
10	NW_001838325	Homo sapiens chromosome 16 genomic scaffold, alternate assembly HuRef SCAF1103279187624, whole genome shotgun sequence	2014/02/08	1862308
11	NM_001126102	Homo sapiens haptoglobin (HP), transcript variant 2, mRNA	2014/01/27	1284
12	NM_005143	Homo sapiens haptoglobin (HP), transcript variant 1, mRNA	2014/01/27	1461
13	NM_00961	Homo sapiens prostacyclin I2 (prostacyclin) synthase (PTGIS), mRNA	2014/01/26	5603
14	NM_020995	Homo sapiens haptoglobin-related protein (HPR), mRNA	2014/01/26	1245
15	NM_0040498	Homo sapiens beta-2-microglobulin (B2M), 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_090699	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_030911	Homo sapiens haptoglobin-related protein (HPR), RefSeqGene on chromosome 16	2013/11/20	21021
21	NC_002952	Staphylococcus aureus subsp. aureus MRSA52 chromosome, complete genome	2013/06/10	2902619
22	NC_003116	Neisseria meningitidis Z2491 chromosome, complete genome	2013/06/10	2184406
23	NZ_ANCJ01000014	Staphylococcus aureus CN79 contig14, whole genome shotgun sequence	2013/05/17	132878
24	NZ_AGRQ01000284	Neisseria meningitidis NM23 ntmf6c 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	JA899727	Sequence 6 from Patent EP2545191	2013/03/27	1461
27	ANCJ01000014	Staphylococcus aureus CN79 contig14, whole genome shotgun sequence	2012/11/19	132878
28	CM000267	Homo sapiens chromosome 16, whole genome shotgun sequence	2009/06/29	75226909
29	CH471166	Homo sapiens 21100065831892 genomic scaffold, whole genome shotgun sequence	2010/02/04	3196546
30	AGR010000284	Neisseria meningitidis NM23 ntmf6c contig283, whole genome shotgun sequence	2012/01/13	3824
31	BC121125	Homo sapiens haptoglobin, mRNA (cDNA clone MGC-150438 IMAGE40120683), complete cds	2006/10/04	1365
32	BC121124	Homo sapiens haptoglobin, mRNA (cDNA clone MGC-150437 IMAGE40120682), complete cds	2006/10/06	1188
33	BC107587	Homo sapiens haptoglobin, mRNA (cDNA clone MGC-111141 IMAGE4734589), complete cds	2005/12/09	1270
34	BC070289	Homo sapiens haptoglobin, mRNA (cDNA clone MGC-88296 IMAGE4716454), complete cds	2004/05/18	1138
35	BC056801	Homo sapiens haptoglobin, mRNA (cDNA clone MGC-61957 IMAGE4723084), complete cds	2006/10/08	1207
36	BC017862	Homo sapiens haptoglobin, mRNA (cDNA clone MGC-22614 IMAGE4700637), complete cds	2006/10/27	1232

Download and Open Download and Save Open at NCBI Total number of hits: 84 more...

- 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

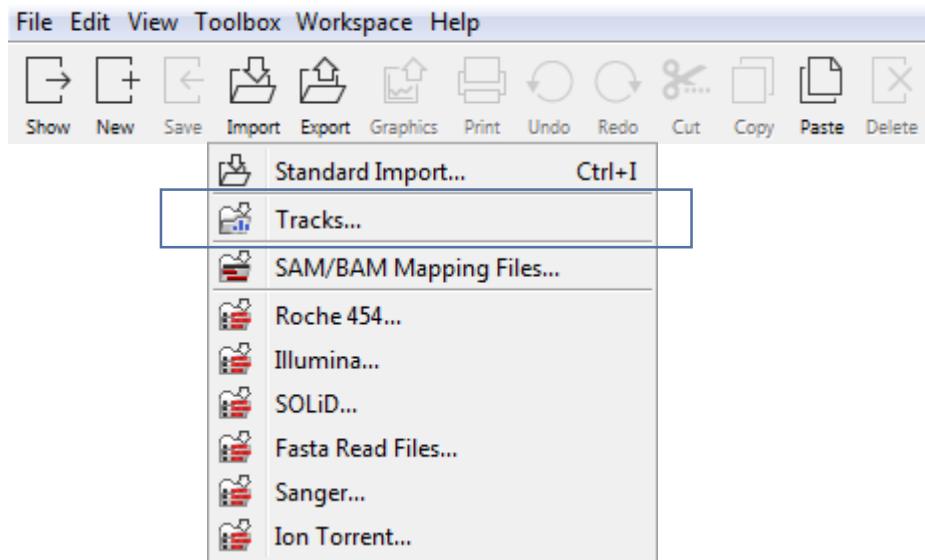
アノテーションインポート

- 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 damp
 - COSMIC variation database

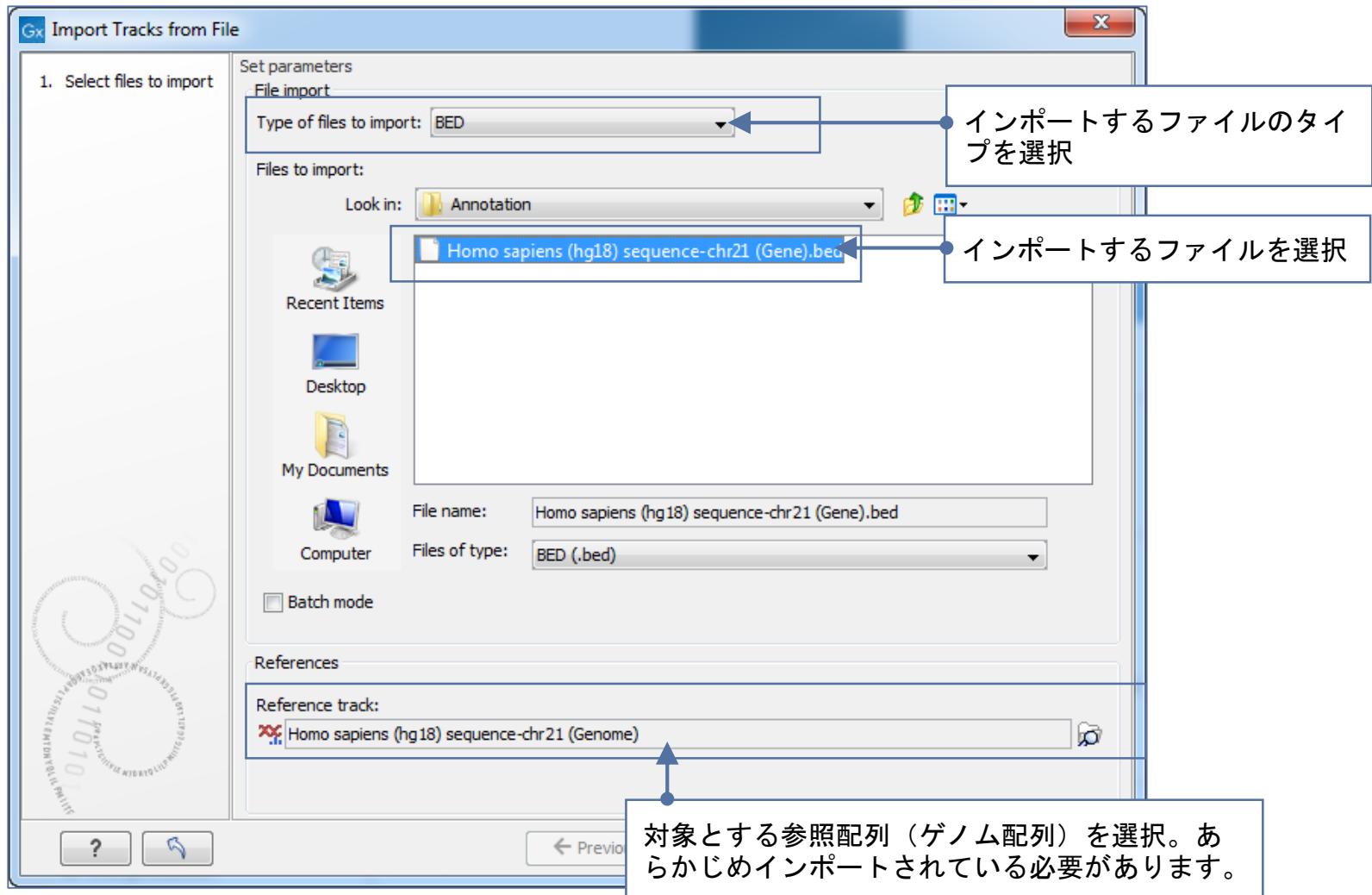
※変異のデータについても、アノテーションとして自分の変異へアノテーションとして情報の追加や比較ができるため、アノテーションのインポート可能フォーマットに含めています。

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



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くらいか、その後のトリミングや、PCR Duplicate の状況などを確認するためにレポートを作成。

トリミング: Trim Sequences

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

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

クオリティスコア

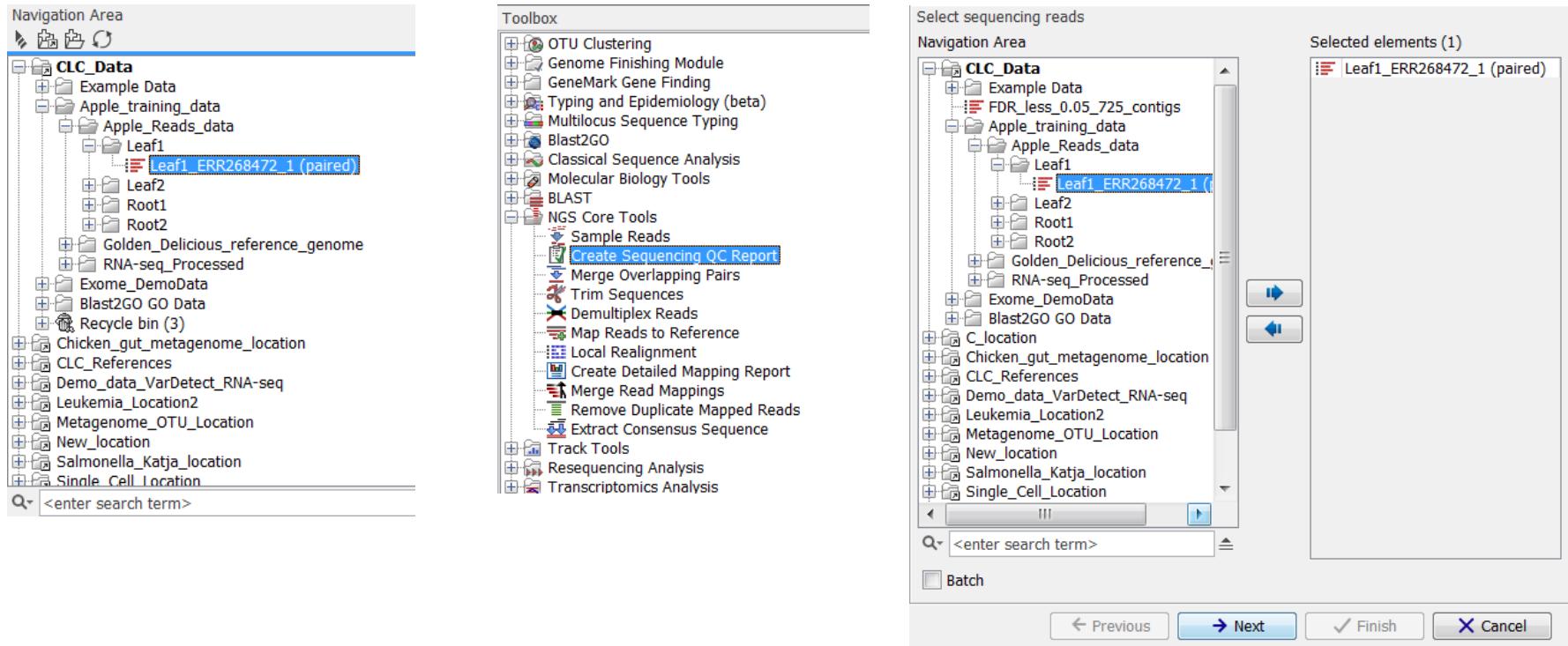
シークエンサーから出てきたリードは、各塩基ごとにエラーの確率の値を持っている。

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

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

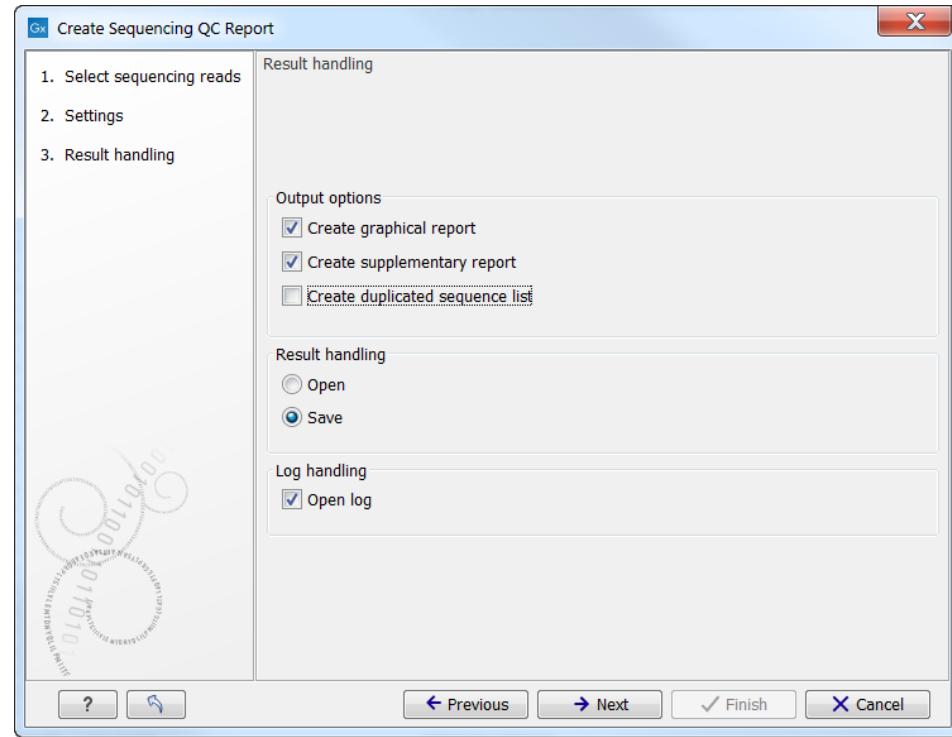
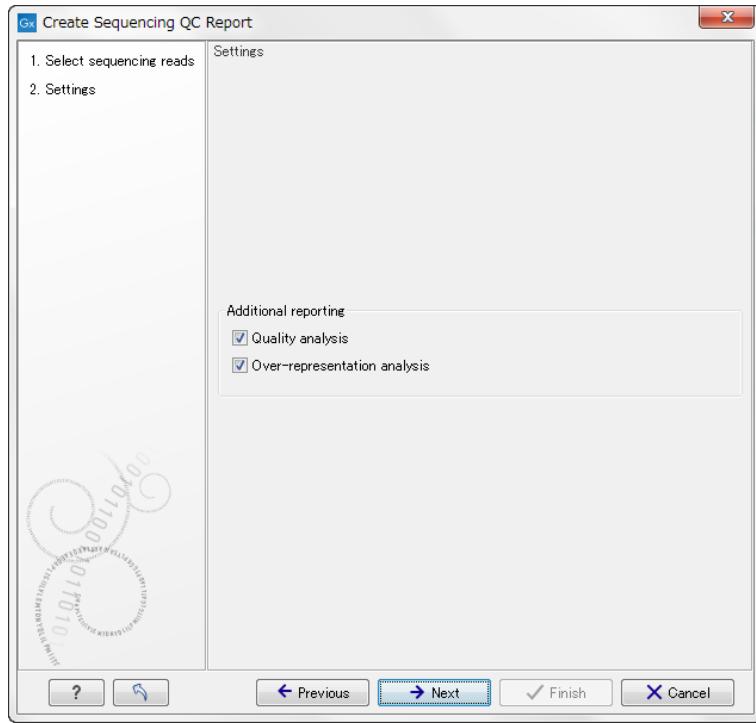
Phred Score	Error の確率	Base call の精度
10	1/10	90%
20	1/100	99%
30	1/1,000	99.9%
40	1/10,000	99.99%
50	1/100,000	99.999%
60	1/1,000,000	99.9999%

QCレポート作成 : Create Sequencing QC Report



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

QCレポート作成 : Create Sequencing QC Report



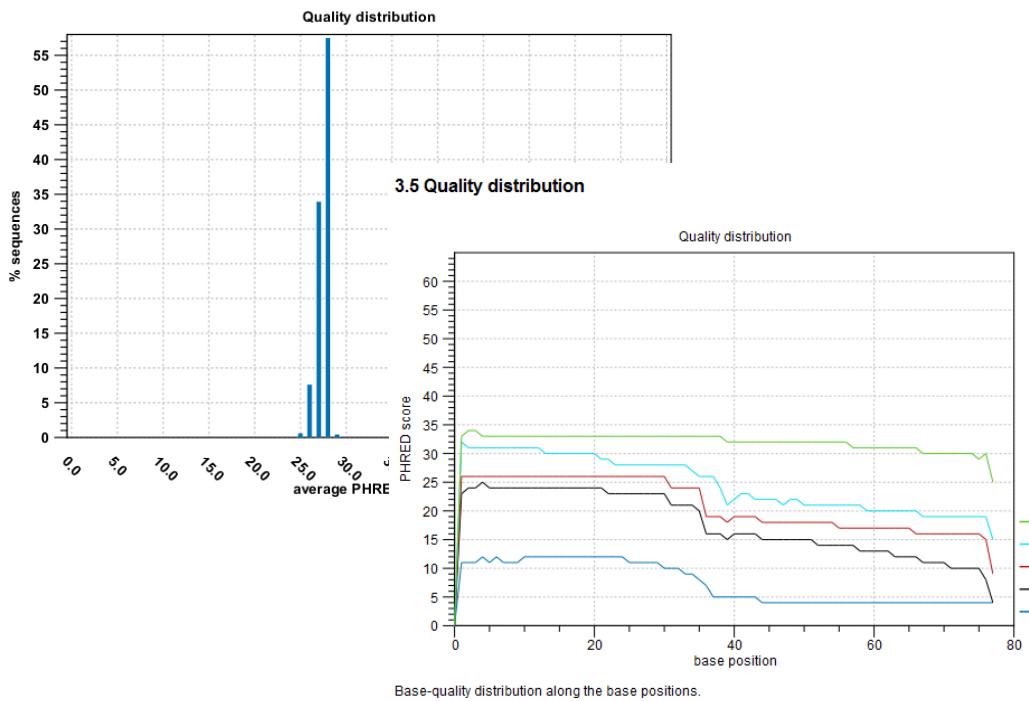
- Quality analysis: クオリティスコアに関する解析。
- Over-representations analysis: 過度に現れているような塩基配列などの解析。

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

QCレポート作成 : Create Sequencing QC Report

- Leaf1_ERR268472_1 (paired)
- Leaf1_ERR268472_1 (paired) - graphical QC report
- Leaf1_ERR268472_1 (paired) - supplementary QC report

2.4 Quality distribution



► 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

- Graphical Report はグラフでのレポートです。
- Supplementary QC Report は、Graphical Report の数字版となり、エクスポートして作図に利用可能です。

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 の低い配列が連續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

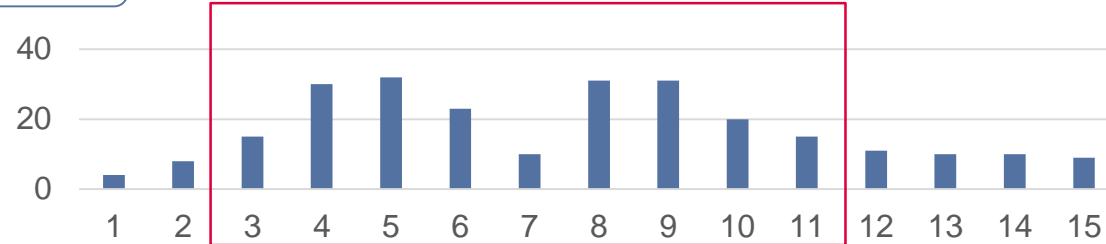
リード配列	G	C	C	C	A	T	G	T	T	C	G	A	T	G	C
Phred score	4	8	15	30	32	23	10	31	31	20	15	11	10	10	9
p値	0.40	0.16	0.03	0.00	0.00	0.01	0.10	0.00	0.00	0.01	0.03	0.08	0.10	0.10	0.13
Limit - p値 (D)	-0.35	-0.11	0.02	0.05	0.05	0.04	-0.05	0.05	0.05	0.04	0.02	-0.03	-0.05	-0.05	-0.08
(D)の累積和	0.00	0.00	0.02	0.07	0.12	0.16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

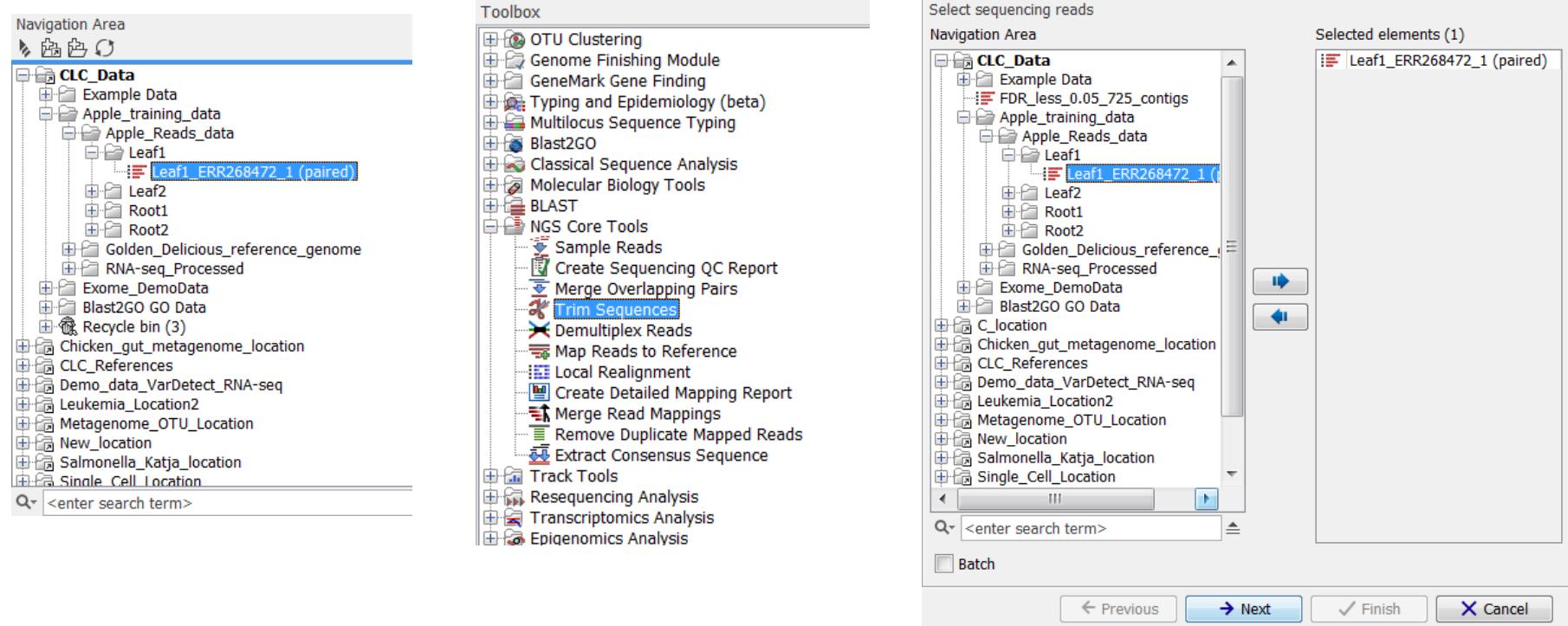
終了点：
累積和が最大を
示す塩基

Phred score の棒グラフ



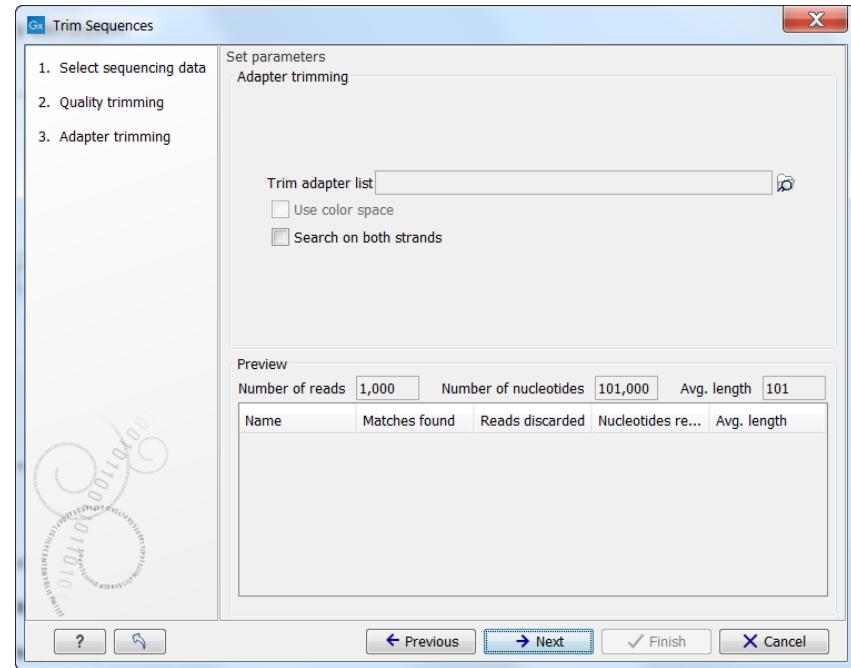
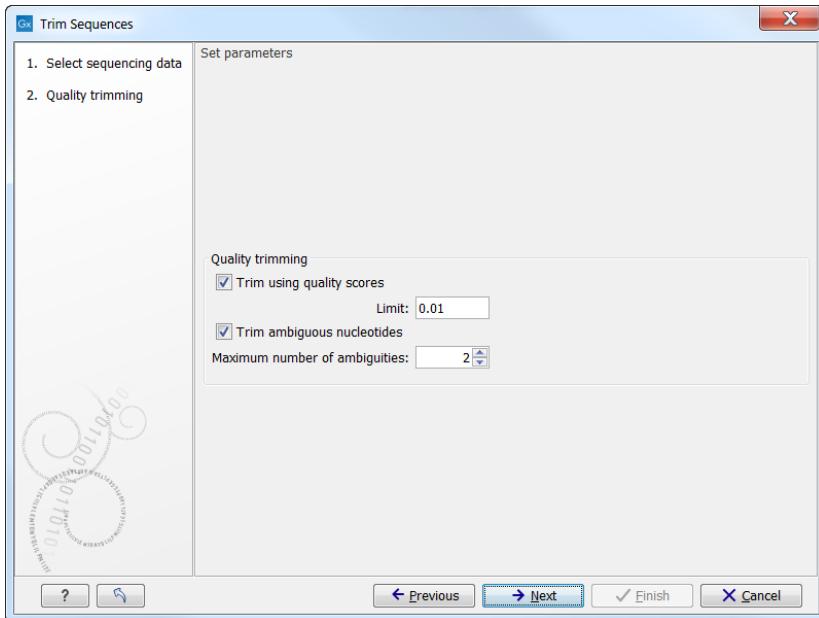
グラフより、ある程度クオリティが高くなっている場所からリードを使い、クオリティが連続して悪くなっている箇所からリードをトリムしていることがわかる。
※途中、1塩基のみクオリティが低いような場合は、必ずしもトリムされない。これはできるだけリードを長く保とうとするため。

トリミング



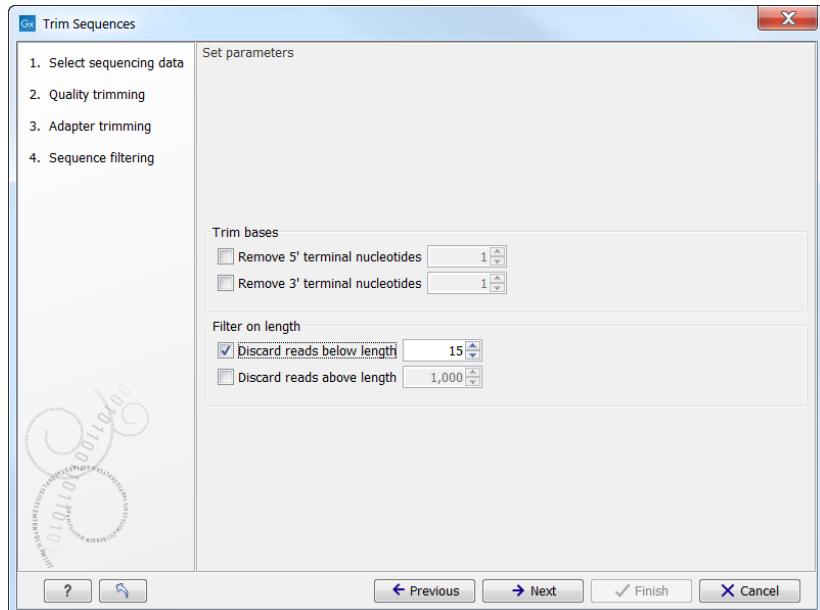
- Navigation Areaから使用するデータを選択。
- Toolboxから Trim Sequences を選択、ダブルクリック。
- ウィザードが起動し、選択したデータが選ばれていることを確認。

トリミング



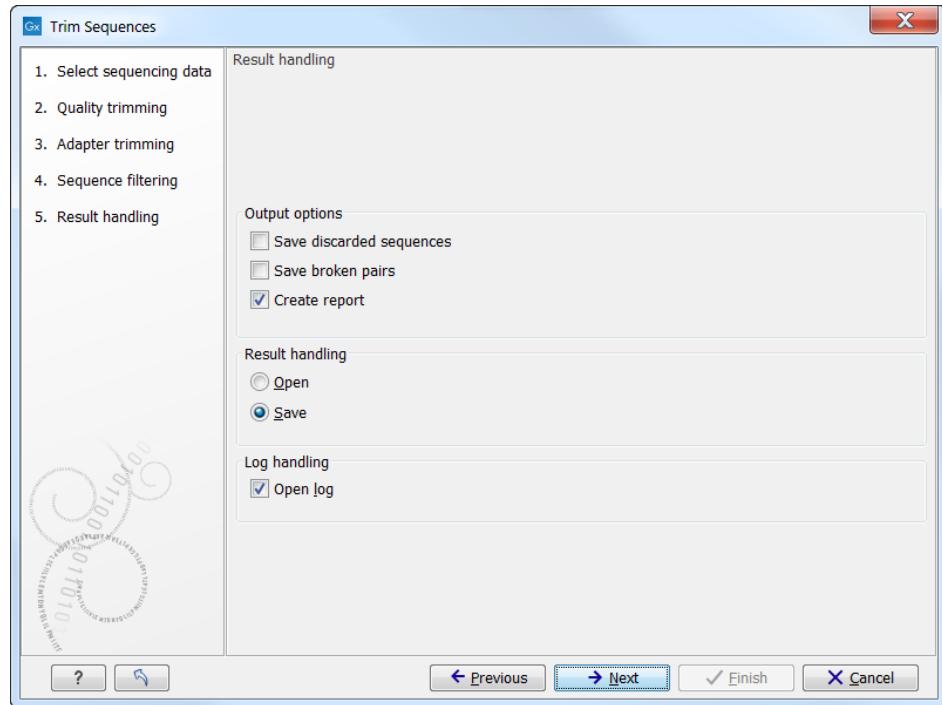
- Trim using quality scores : トリミングに使用するLimitパラメータを決定
- Trim ambiguous nucleotides : N表示される塩基について、最大何塩基まで保持させるか。

- 今回はアダプターは設定なし。



Trim bases

- 5末、3末の塩基数を指定してカット
- Filter on length
- 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



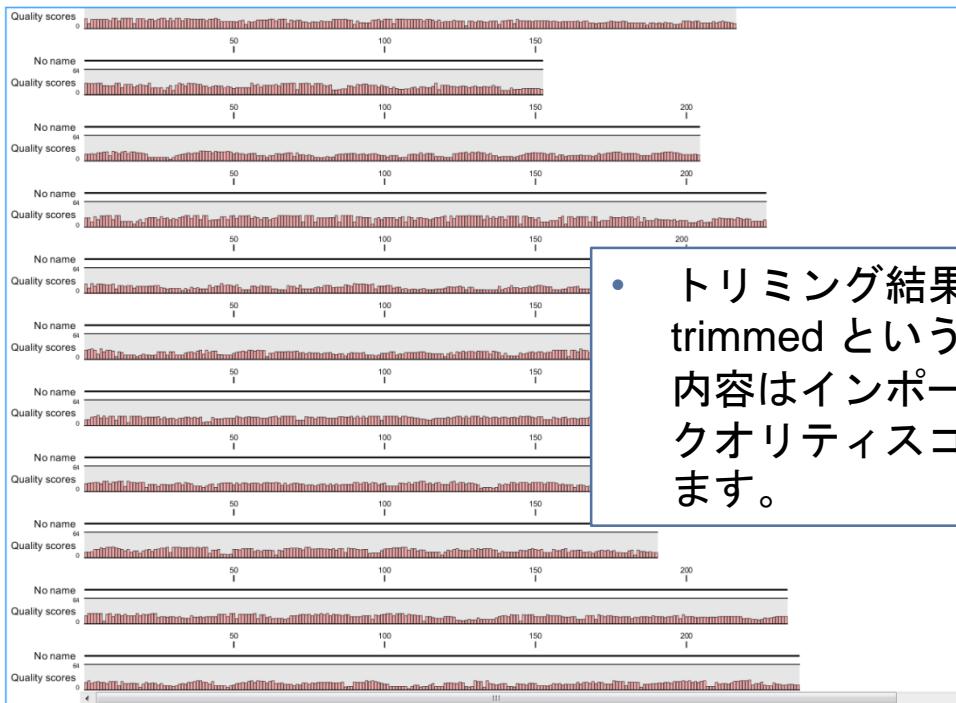
- レポートの作成にチェック。

トリミング結果

結果

- Leaf1_ERR268472_1 (paired)
- Leaf1_ERR268472_1 (paired) - graphical QC report
- Leaf1_ERR268472_1 (paired) - supplementary QC report
- Leaf1_ERR268472_1 (paired) trimmed (paired)
- Leaf1_ERR268472_1 (paired) report

- トリミング後は、トリムされたリードと、レポートを作成を選択した場合は、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

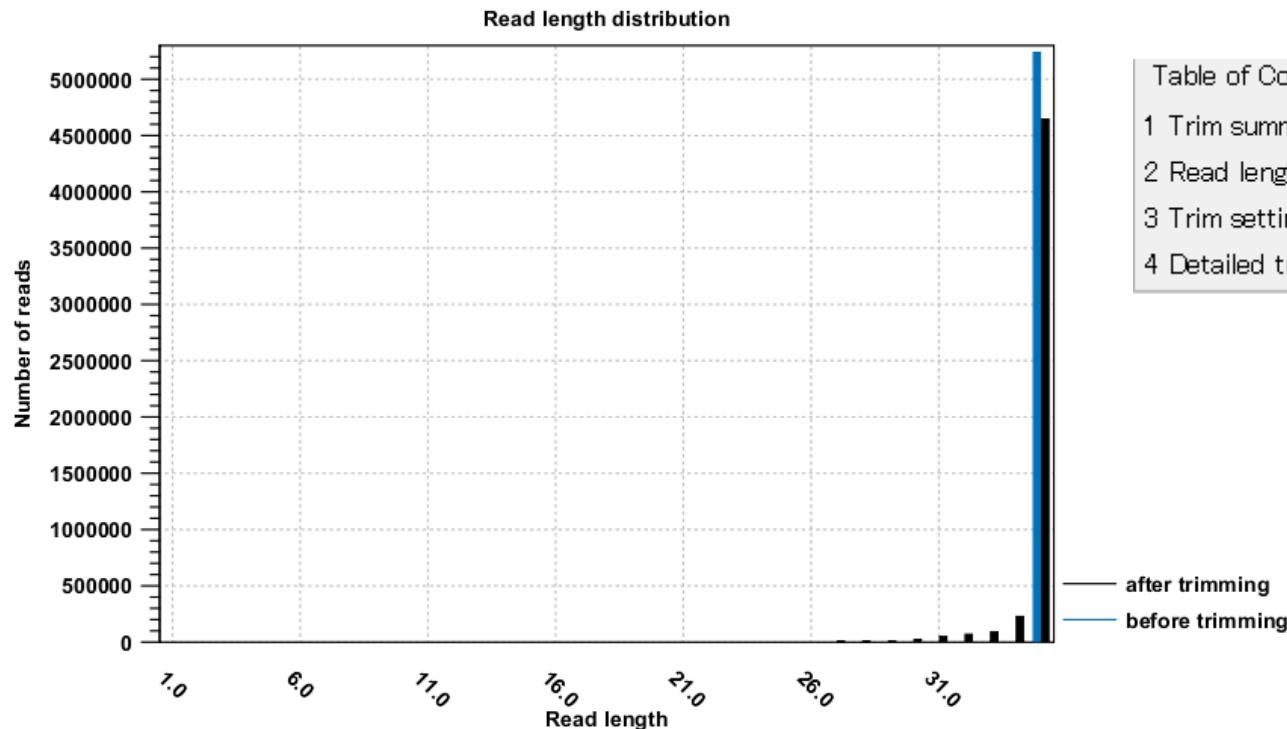


Table of Contents

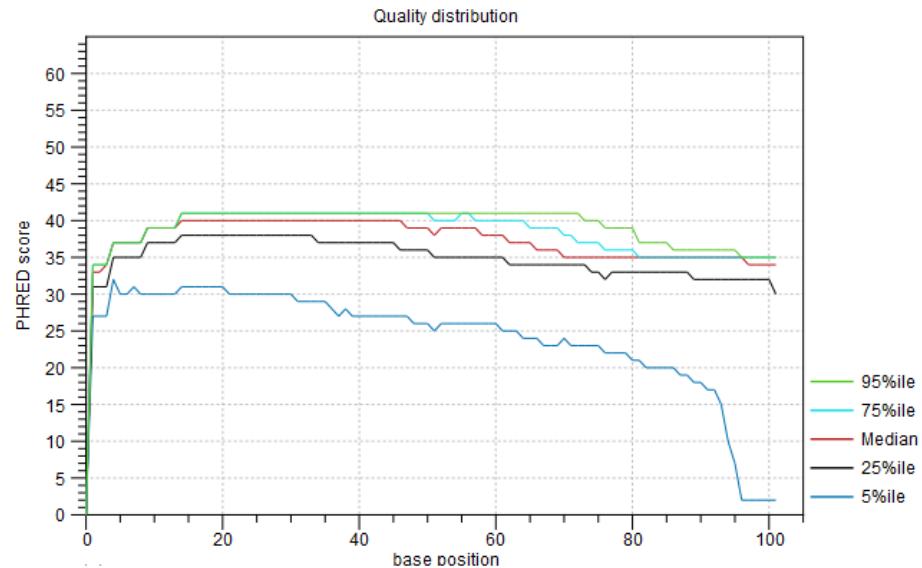
- 1 Trim summary
- 2 Read length before / after trimming
- 3 Trim settings
- 4 Detailed trim results

エクササイズ

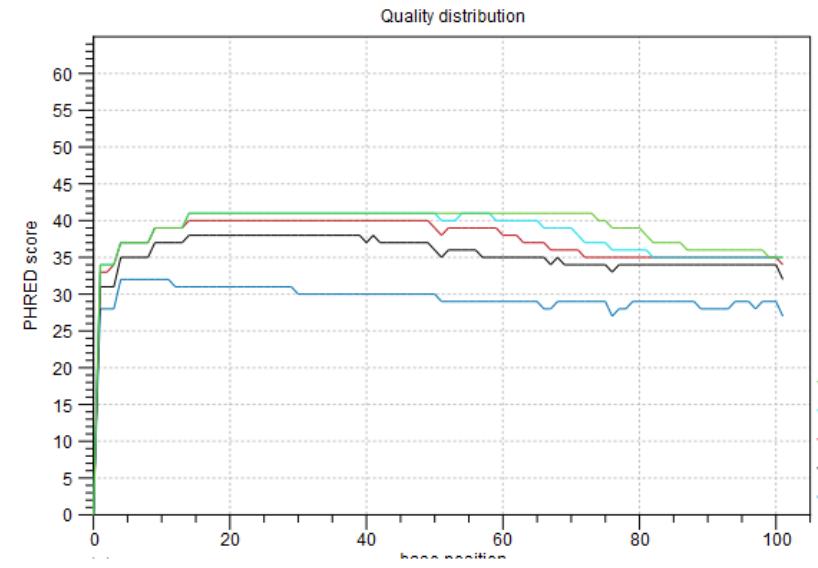
- トリミング後のデータでレポートを作成してみましょう！

Before

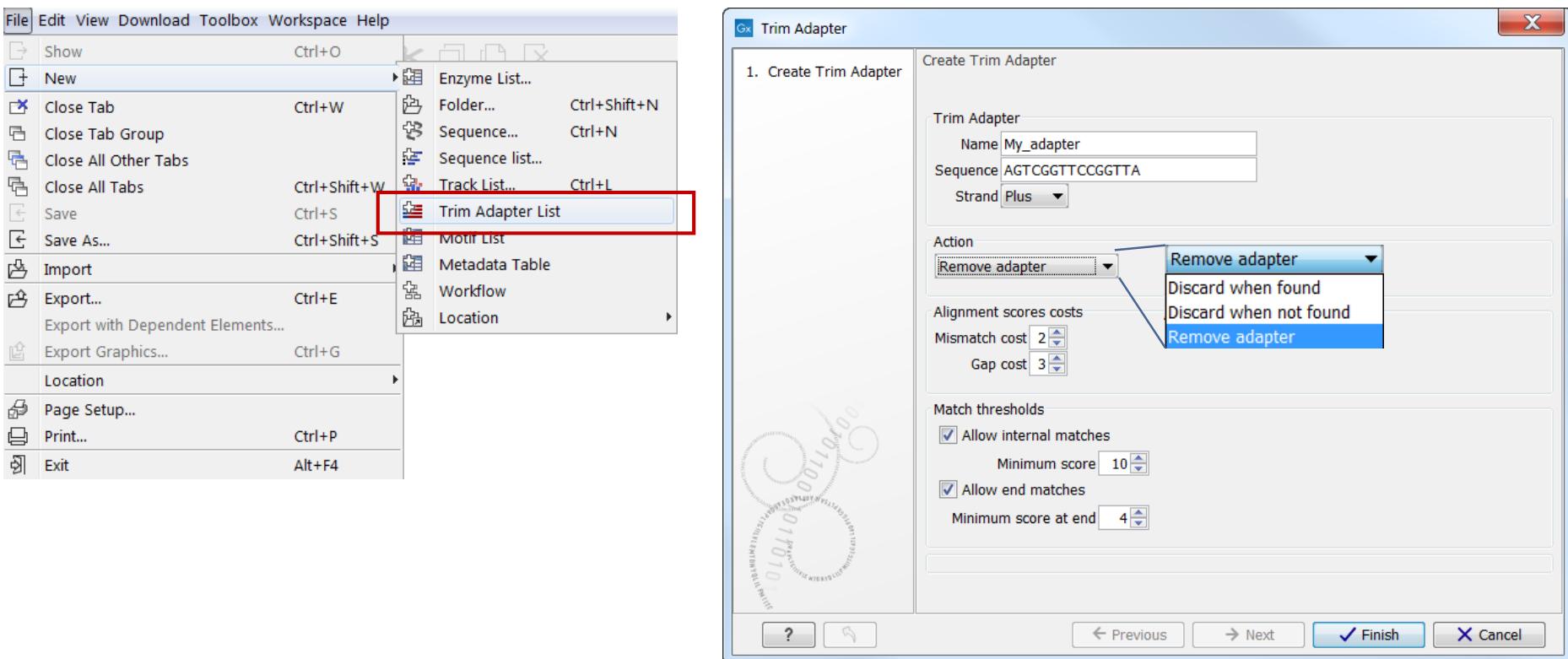
3.5 Quality distribution

**After**

3.5 Quality distribution



アダプターリストの作成

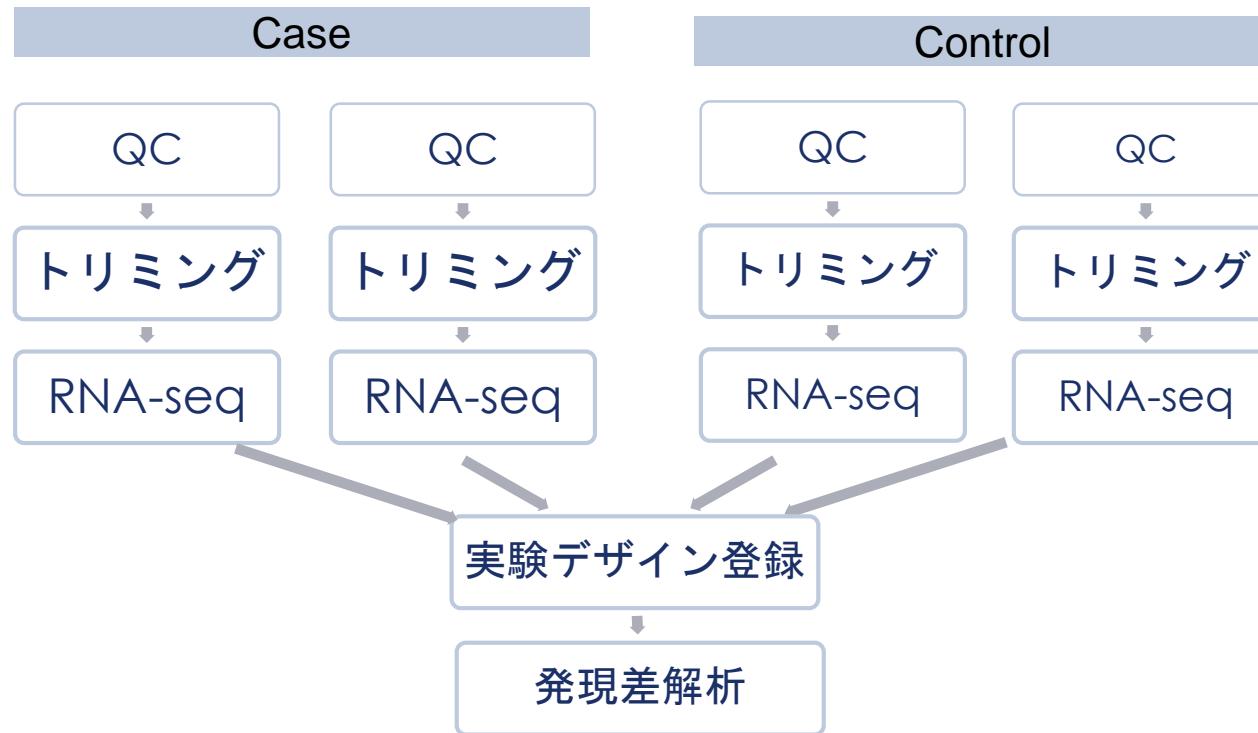


- 作成されたアダプターリストは、Trimmingツールの中で指定することができます。

RNA-seqと発現差解析

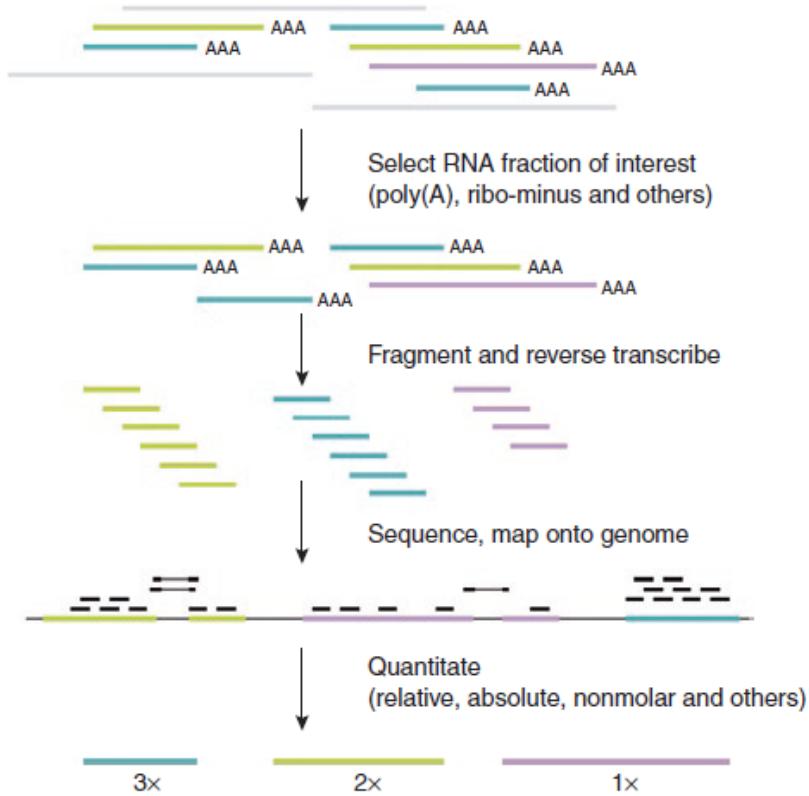
株式会社キアゲン
グローバルリフォマティクス ソリューション&サポート
アプライドアドバンストゲノミクス
宮本 真理
mari.miymoto@qiagen.com

実験デザイン



RNA-seqマッピング原理

RNA-seq : 原理



RNA抽出

cDNA作成

Gene, Transcriptへ
マッピング

Pepke, S.; Wold, B. & Mortazavi, A.
Computation for ChIP-seq and RNA-seq studies
Nature methods, Nature Publishing Group, 2009, 6, S22-S32

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

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

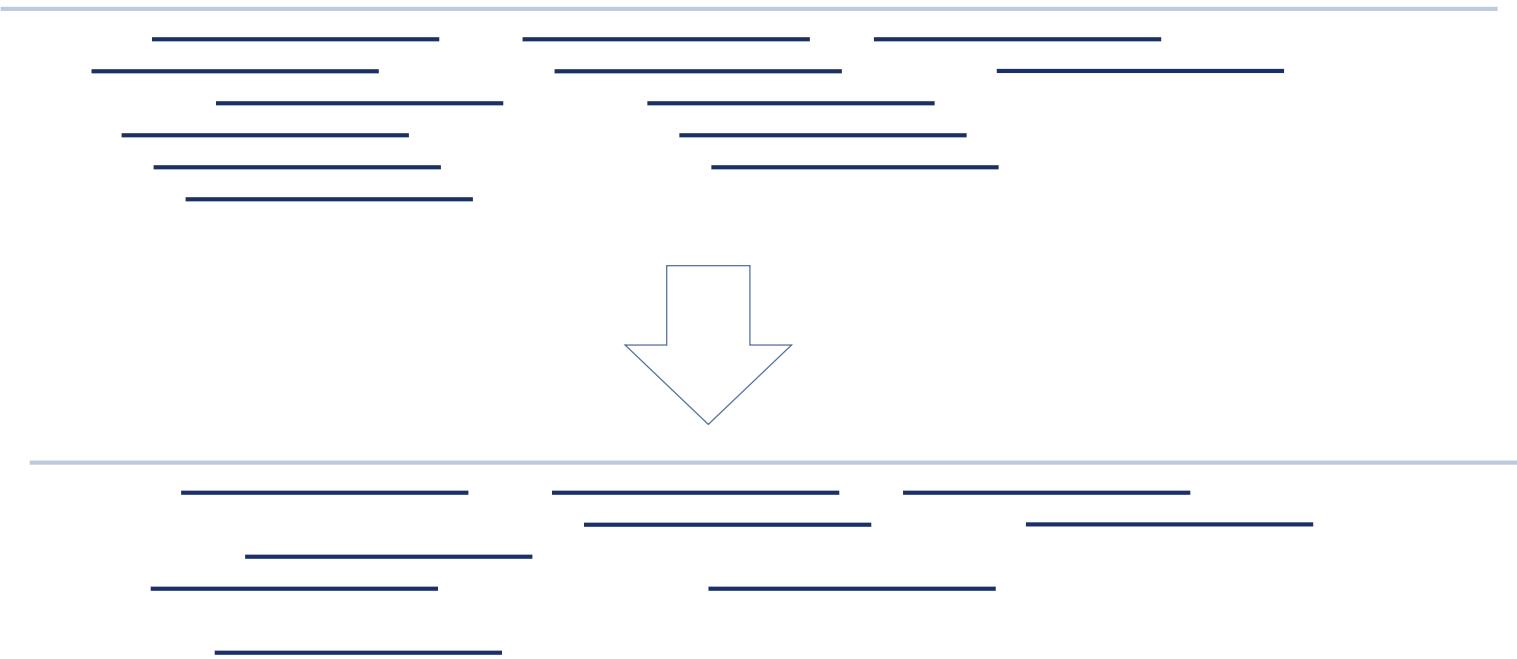
アライメントスコア = 16

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。



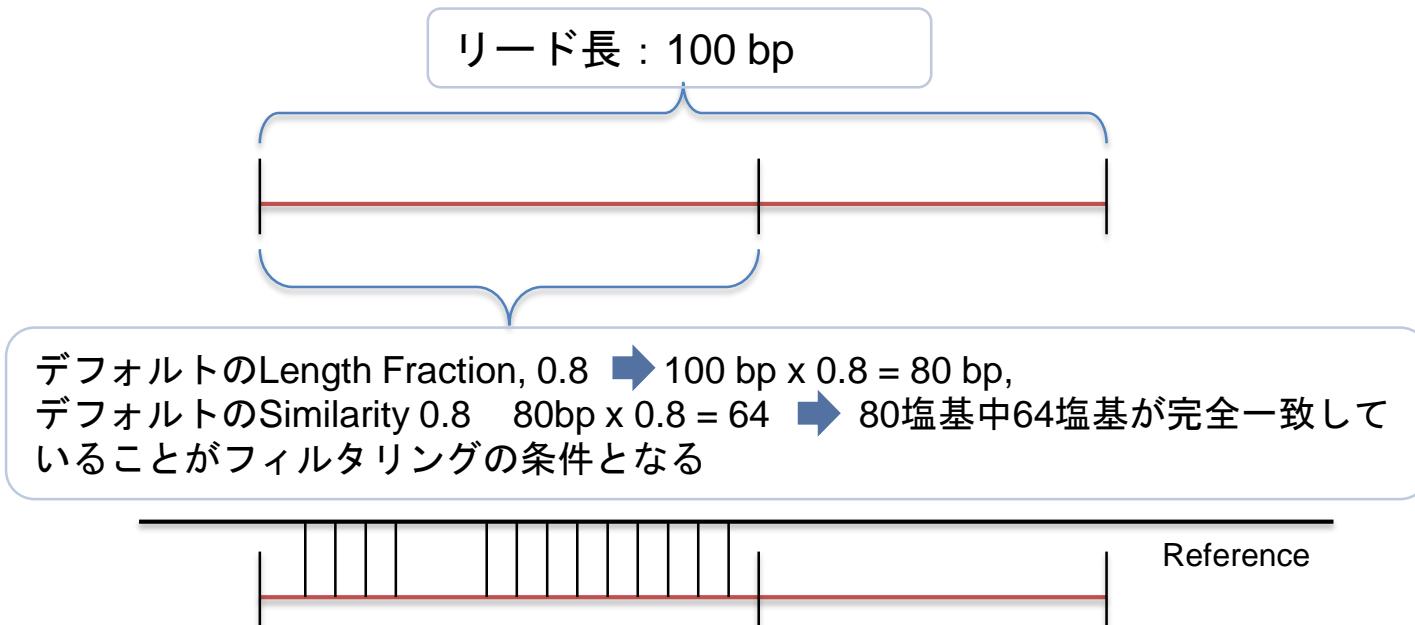
フィルタリング原理

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。

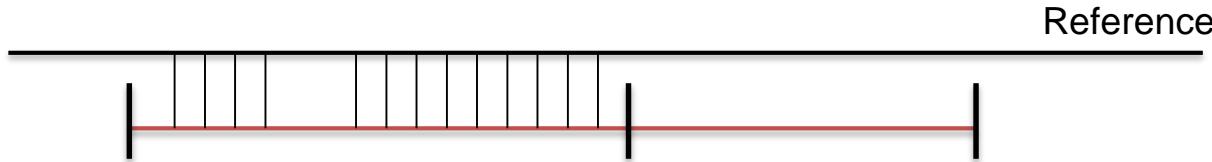
Length Fraction と Similarity は 2 つのパラメータの組み合わせで使用されます。

Length fraction: フィルターをかける際に、考慮する長さ

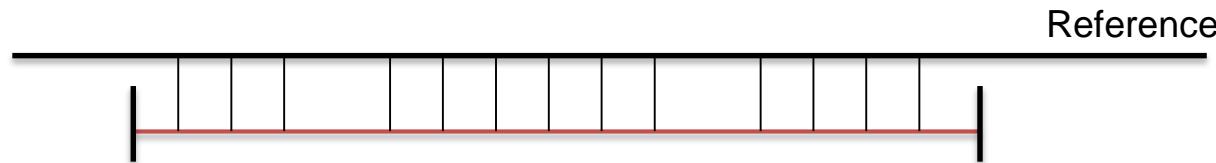
Similarity: Length Fraction で指定した長さのうち、どの程度類似しているものを残すか。



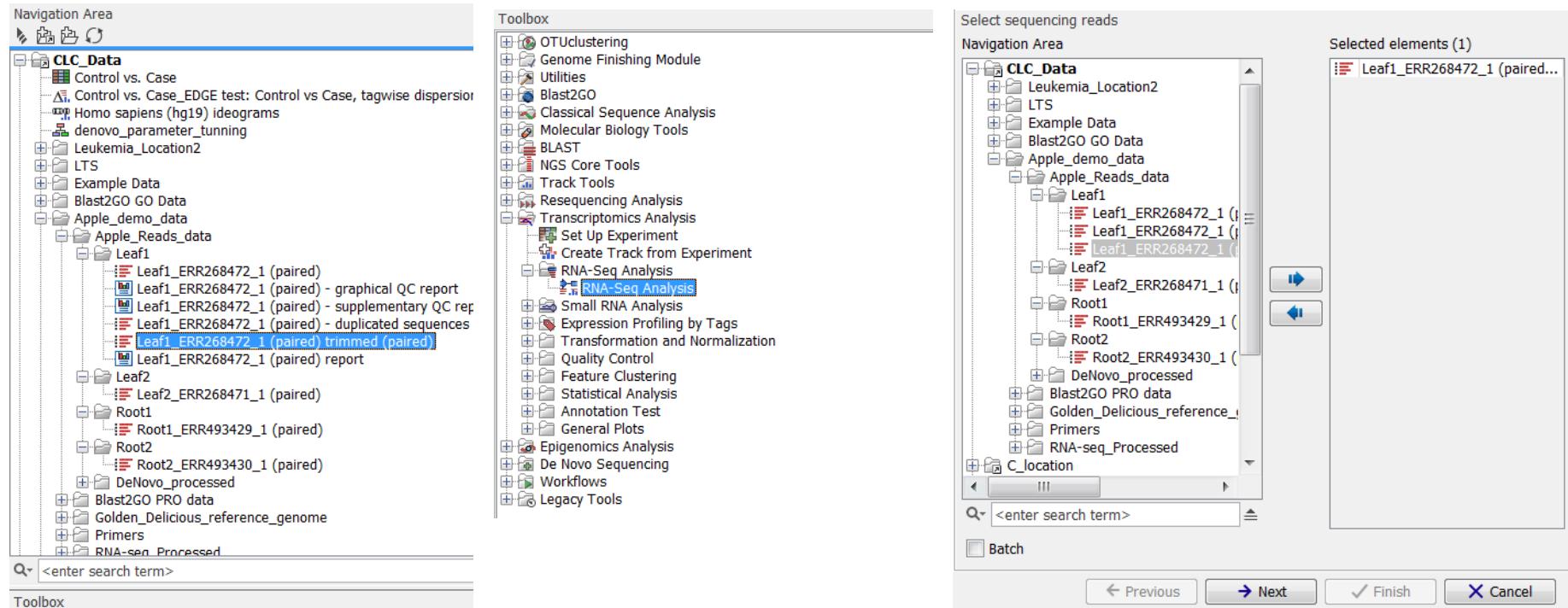
2つのパラメータを使う理由



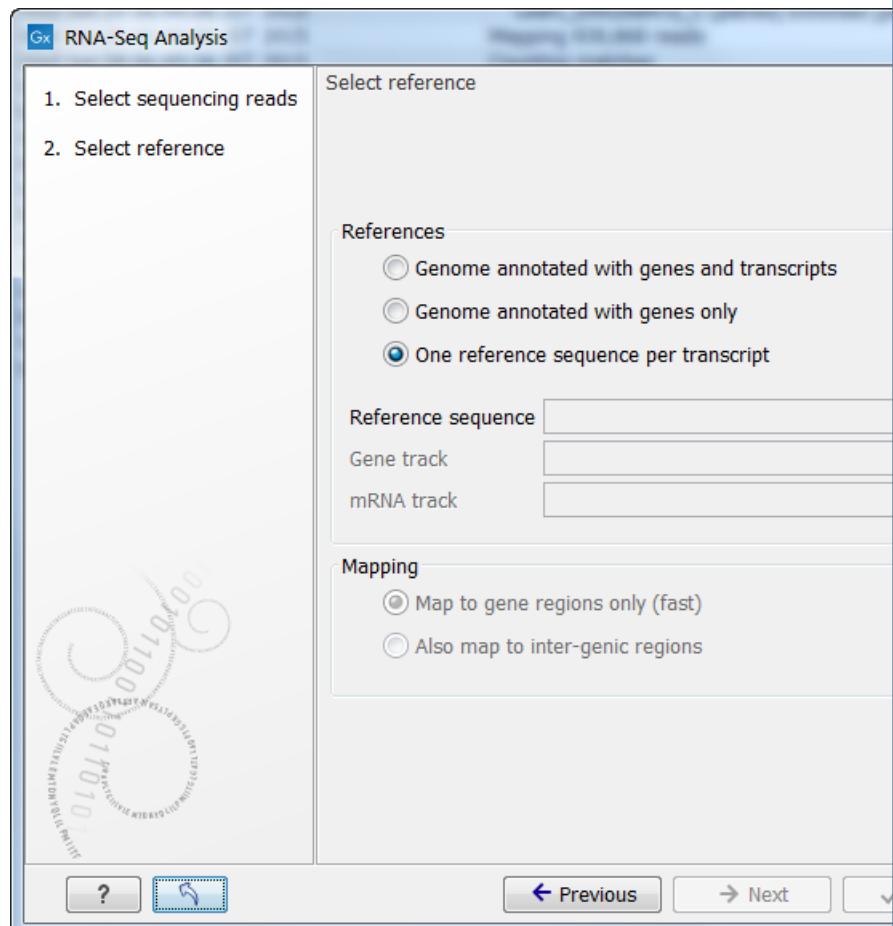
- リードの一部は似ているけれども、大きな挿入や、欠失によりリードの一部が参照配列と一致しない可能性がある場合
- トリミングが完全にできなかったクオリティの低い配列が末端部にある場合
(Length Fraction を小さくすることで、リードの一部に限定してアライメントの類似度を設定できる)



- 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。

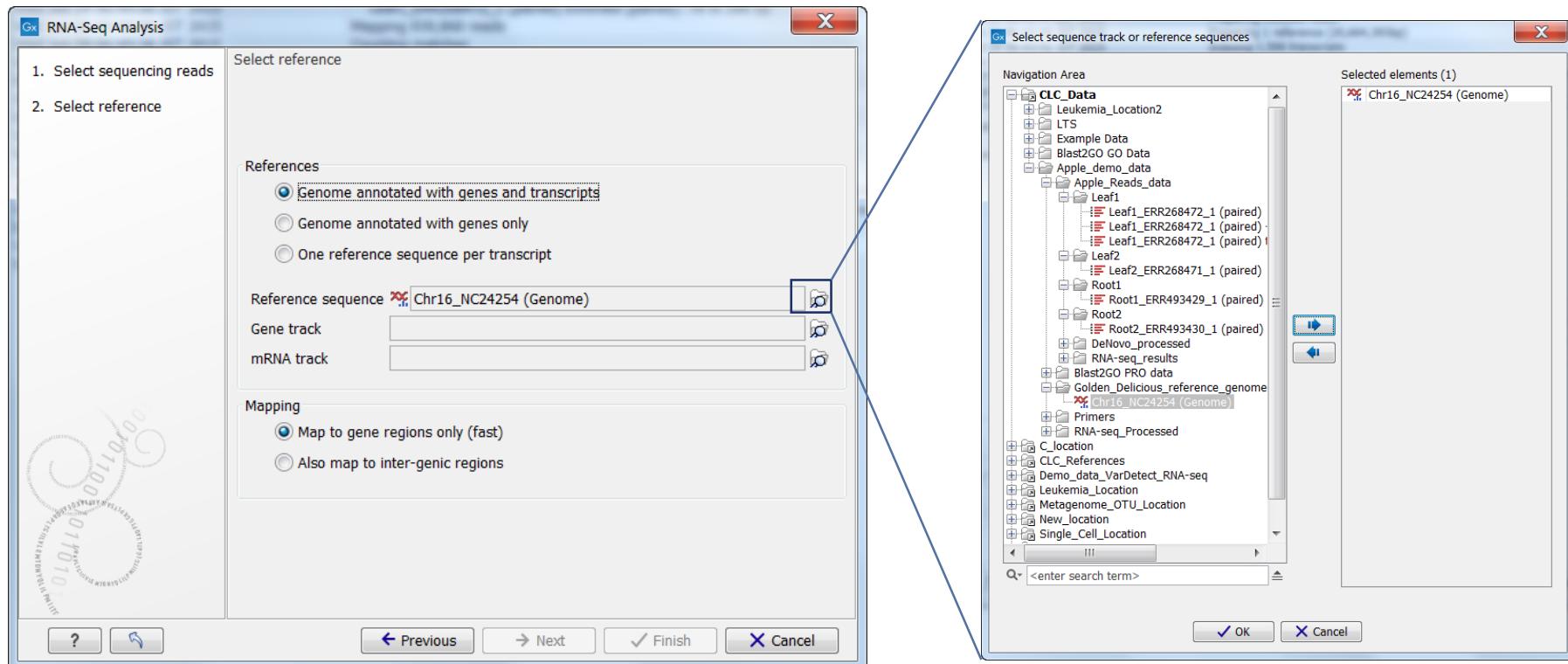


References

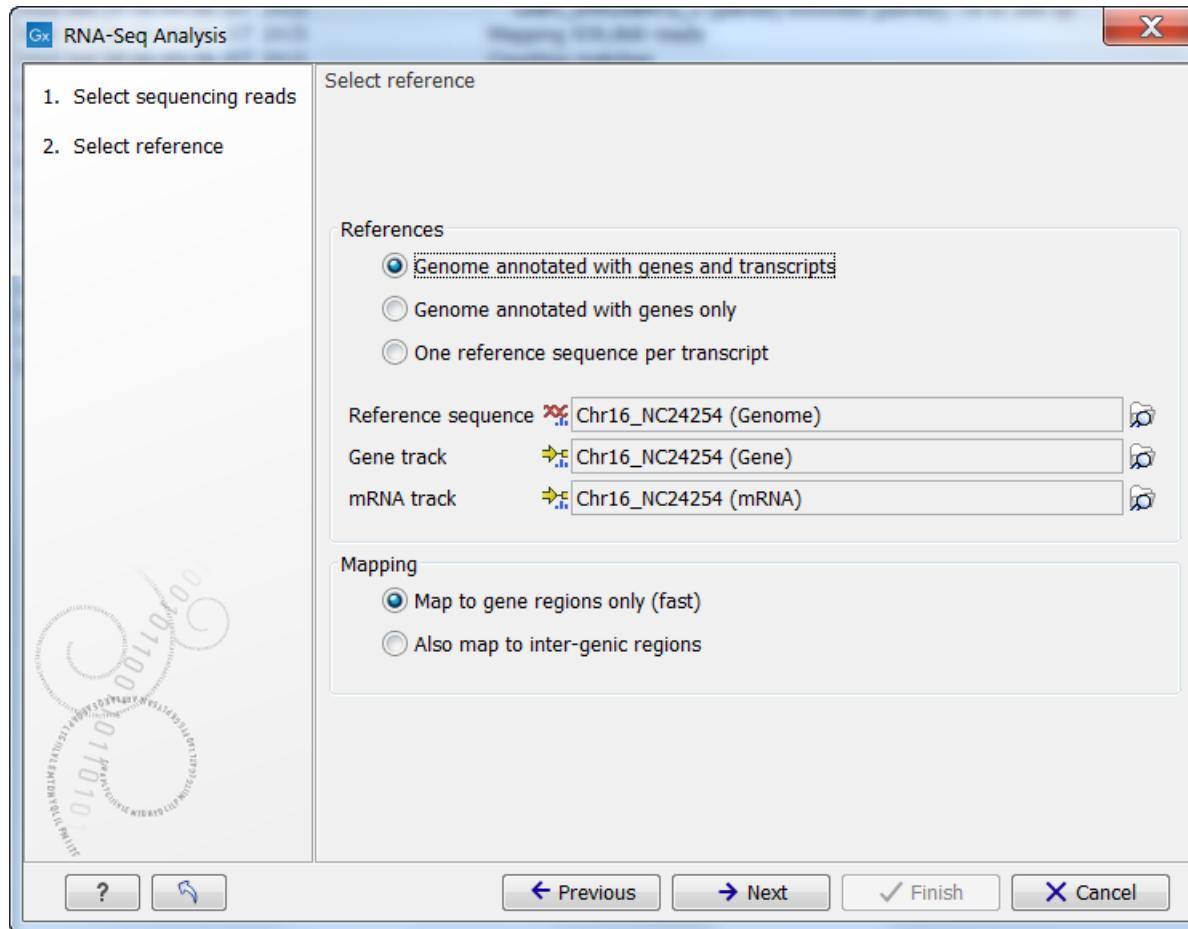
- Genome annotated with genes and transcripts: ゲノムに Gene と Transcript のアノテーションがある場合 (Eukaryote)
- Genome annotated with gene only: ゲノムに Gene のアノテーションのみある場合 (Prokaryote)
- One reference sequence per transcript: 1 つの配列を 1 つの Transcript とみなす場合
- Reference sequence, gene, mRNA の track のうち必要なものを選択 (上記 Reference の選択により必要なものが異なる)

Mapping

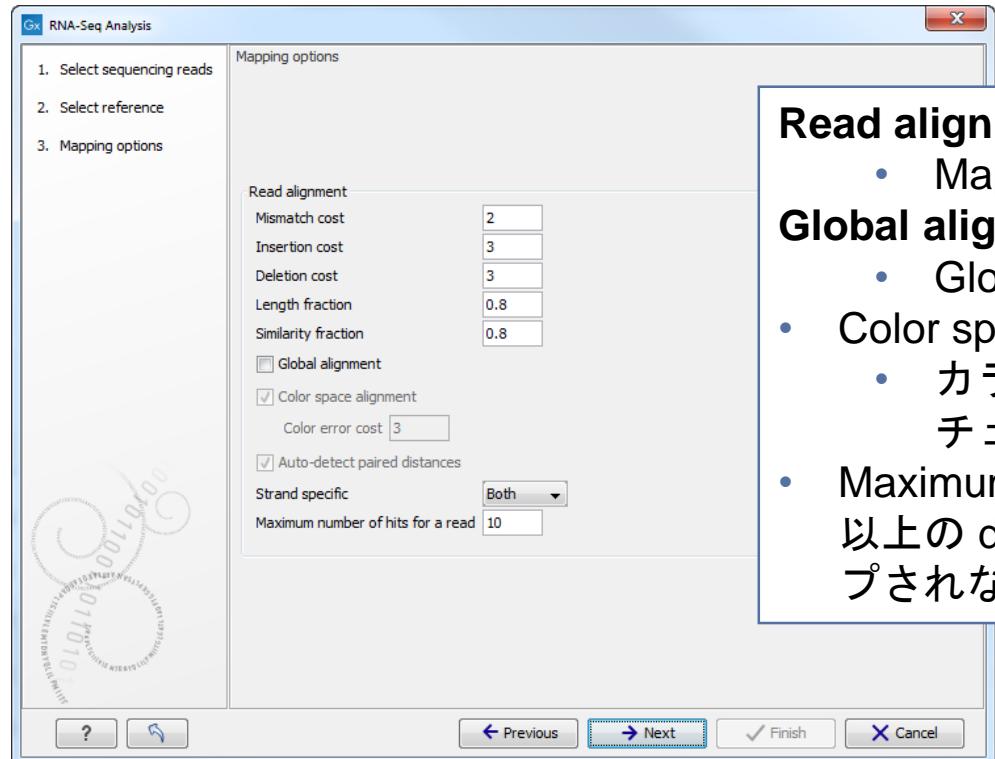
- Gene regions のみにマップ
- Inter-genic regions へもマップ



- 今回はゴールデンデリシャスのゲノムへマップするため、
Genomeと関連するアノテーションをそれぞれ選択します。



- Genome, Gene, mRNA を選択すると、上記のようになります。

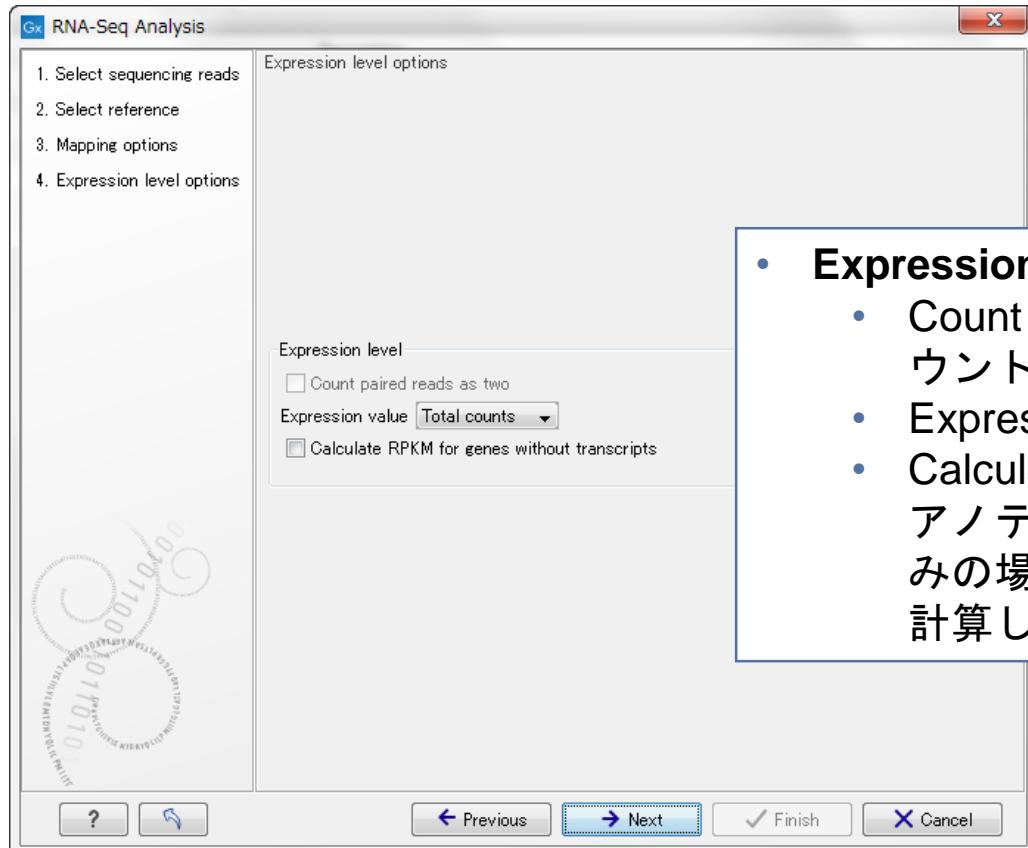


Read alignment

- Mapping の原理参照

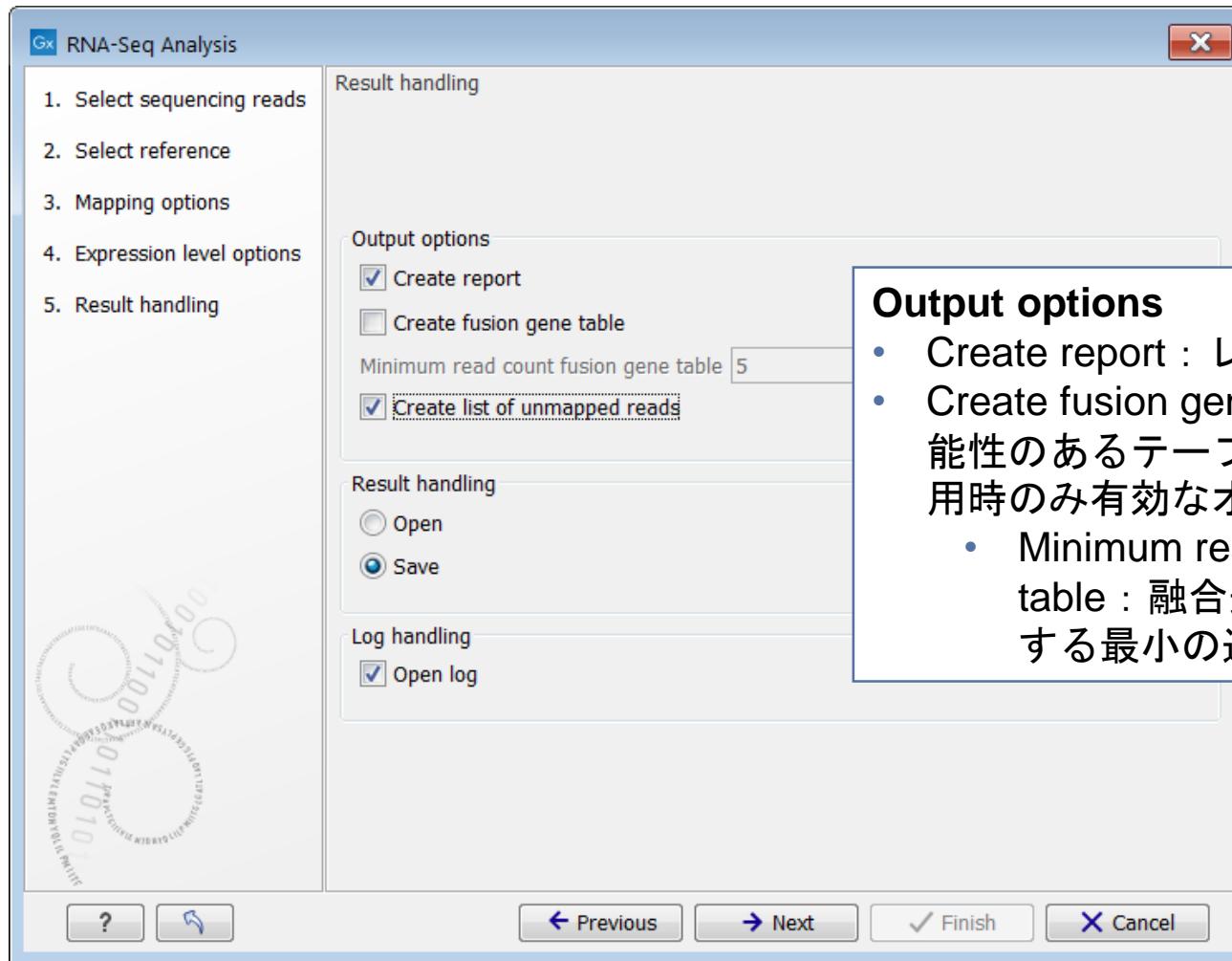
Global alignment

- Global alignment を行う場合にチェック
- Color space alignment
 - カラースペースを使う場合 (SOLiD) にチェック
- Maximum number of hits for a read: 指定した数以上の distinct places にマッチした場合、マップされない



- **Expression level:**

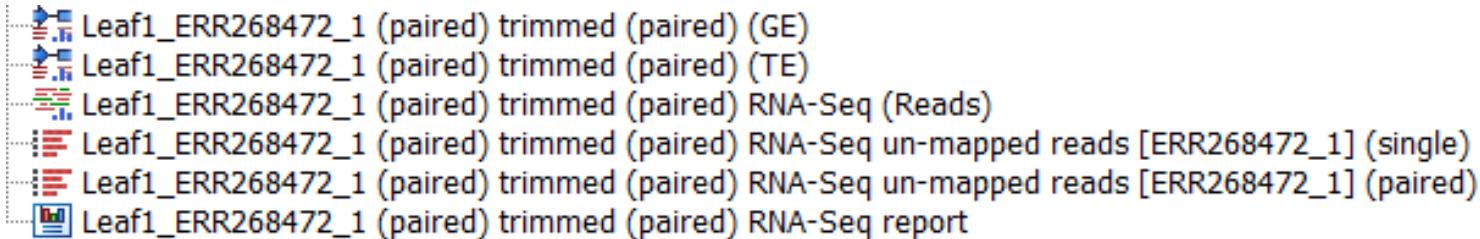
- Count paired reads as two: ペアを2リードとカウントしたい場合
- Expression value : 発現量に何を指定するか
- Calculate RPKM for gene without transcripts: アノテーションとしてmRNAがなく、遺伝子のみの場合。この場合、遺伝子の全長でRPKMを計算します。



Output options

- Create report : レポートの作成
- Create fusion gene table : 融合遺伝子の可能性のあるテーブルの作成（ペアエンド利用時のみ有効なオプション）
 - Minimum read count fusion gene table : 融合遺伝子の可能性があるとする最小の遺伝子

結果



アウトプットで指定をしていないものは作成されないため、ご注意ください。

- ファイル名 (GE) : Gene Expression ト ラック
- ファイル名 (TE) : Transcript Expression ト ラック
- ファイル名 (Reads) : マッピングト ラック
- ファイル名 (single),(paired): マッピングト ラック
- ファイル名 report: レポート

結果

Leaf1_ERR268472_1 (paired) trimmed (paired) (GE)

Leaf1_ERR268472_1 (paired) trimmed (paired) (TE)

Leaf1_ERR268472_1 (paired) trimmed (paired) RNA-Seq (Reads)

Leaf1_ERR268472_1 (paired) trimmed (paired) RNA-Seq un-mapped reads [ERR268472_1] (single)

Leaf1_ERR268472_1 (paired) trimmed (paired) RNA-Seq un-mapped reads [ERR268472_1] (paired)

Leaf1_ERR268472_1 (paired) trimmed (paired) RNA-Seq report

Show column

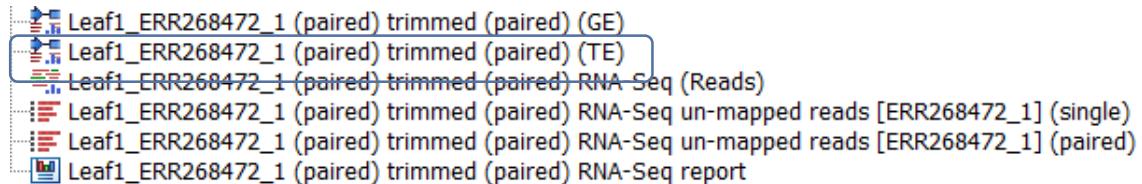
- Name
- Chromosome
- Region
- Expression value
- Gene length
- RPKM
- Unique gene reads
- Total gene reads
- Transcripts annotated
- Detected transcripts
- Exon length
- Exons
- Unique exon reads
- Total exon reads
- Ratio of unique to total (exon reads)
- Unique exon-exon reads
- Total exon-exon reads
- Unique intron reads
- Total intron reads
- Ratio of intron to total gene reads

Select All

Deselect All

Name	Chromosome	Region	Expression value	Gene length	RPKM	Unique ge...	Total gen...	Transcript...	Detected t...	Exon
LOC103403208	NC_024254	complement(1325..16185)	0.00	14861	0.00	188	189	0	0	
LOC103402574	NC_024254	7859..8713	1.00	855	2.81	1	1	1	1	
LOC103403209	NC_024254	16284..19959	0.00	3676	0.00	0	0	1	0	
LOC103402575	NC_024254	21003..21678	147.00	676	587.99	151	151	1	1	
LOC103403210	NC_024254	complement(40627..42248)	0.00	1622	0.00	0	0	1	0	
LOC103402576	NC_024254	complement(45595..46594)	0.00	1000	0.00	0	0	1	0	
LOC103402577	NC_024254	complement(46691..47986)	0.00	1296	0.00	0	0	1	0	
LOC103402578	NC_024254	52542..54115	0.00	1574	0.00	9	9	0	0	
LOC103402579	NC_024254	complement(68342..69268)	729.00	927	1,921.14	731	731	1	1	
LOC103402580	NC_024254	72005..77634	82.00	5630	70.21	88	88	1	1	
LOC103402584	NC_024254	78425..81714	534.00	3290	836.49	539	539	1	1	
LOC103402583	NC_024254	complement(82254..85186)	472.00	2933	674.46	492	492	3	0	
LOC103402582	NC_024254	88331..89524	0.00	1194	0.00	1	1	2	0	
LOC103402581	NC_024254	95621..98626	105.00	3006	142.95	106	106	1	1	
LOC103402586	NC_024254	101868..102642	17.00	775	52.83	18	18	1	1	
LOC103403212	NC_024254	106367..111417	262.00	5051	274.35	271	271	1	1	
LOC103402587	NC_024254	111966..115108	202.00	3143	266.83	207	207	1	1	
LOC103402588	NC_024254	complement(115403..121149)	2,315.00	5747	1,653.43	2327	2327	1	1	
TRNAS-AGA	NC_024254	complement(124960..125041)	0.00	82	0.00	0	0	0	0	
LOC103402589	NC_024254	128752..129756	87.00	1005	234.19	88	88	1	1	
LOC103402590	NC_024254	complement(132452..136510)	207.00	4059	248.77	212	212	5	2	
LOC103402591	NC_024254	141406..144140	71.00	2735	167.26	73	73	1	1	
LOC103402592	NC_024254	complement(144604..146373)	8.00	1770	20.93	8	8	1	1	

結果



Show column

Name
 Chromosome
 Region
 Expression value
 Gene length
 RPKM
 Unique gene reads
 Total gene reads
 Transcripts annotated
 Detected transcripts
 Exon length
 Exons
 Unique exon reads
 Total exon reads
 Ratio of unique to total (exon reads)
 Unique exon-exon reads
 Total exon-exon reads
 Unique intron reads
 Total intron reads
 Ratio of intron to total gene reads

Select All Deselect All

Transcript Expression ツラック

転写因子レベルでの発現量が表示されています。

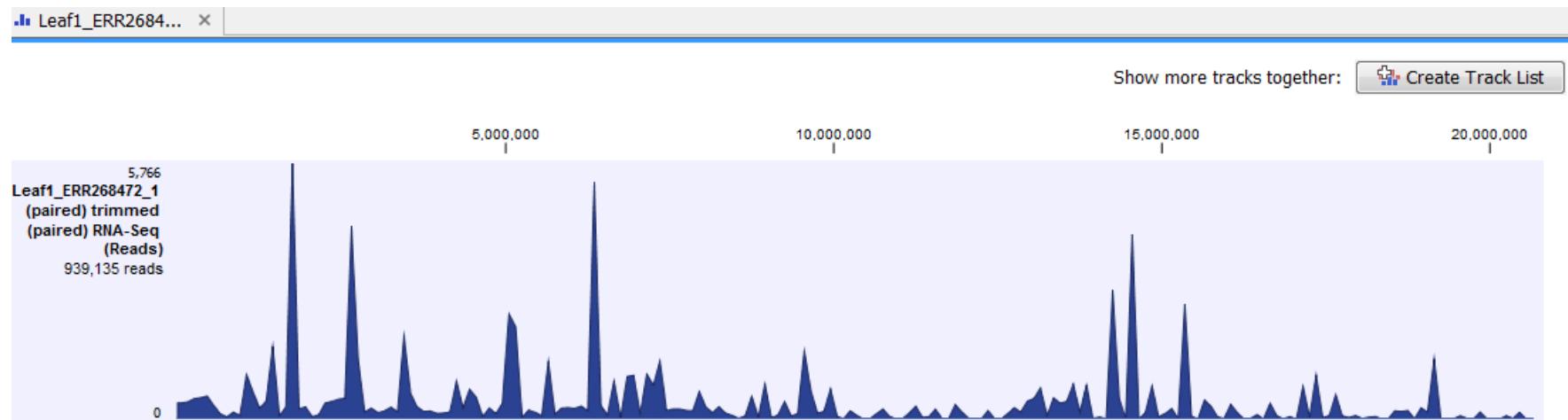
Rows: 1,506 Table view: Genome		
Name	Chromosome	Region
LOC103402574_1	NC_024254	join(7859..8045,8142..8713)
LOC103403209_1	NC_024254	join(16284..16385,19023..19191,19497..19959)
LOC103402575_1	NC_024254	join(21003..21235,21379..21678)
LOC103403210_1	NC_024254	complement(join(40627..40942,42032..42248))
LOC103402576_1	NC_024254	complement(45595..46594)
LOC103402577_1	NC_024254	complement(join(46691..47189,47712..47986))
LOC103402579_1	NC_024254	complement(join(68342..68847,68966..69268))
LOC103402580_1	NC_024254	join(72005..72526,73122..73191,73410..73477,73613..73701,74204..74401,74973..75177,75288..75455,75855..76022,76109..76299,76824..77634); join(78425..78639,78787..78907,79052..79116,79299..79367,79504..79622,80409..80556,80713..80857,80966..81052,81323..81714)
LOC103402584_1	NC_024254	join(82254..82637,82961..83074,83176..83314,83406..83484,83589..83647,83869..84134,84811..85149)
LOC103402583_1	NC_024254	complement(join(82254..82637,82961..83074,83176..83314,83406..83484,83589..83647,83869..84134,85136..85186))
LOC103402583_2	NC_024254	complement(join(82254..82637,82961..83074,83176..83314,83406..83484,83589..83647,83869..84134,85136..85186))
LOC103402583_3	NC_024254	complement(join(82254..82637,82961..83074,83176..83314,83406..83484,83589..83647,83869..84134,85136..85186))
LOC103402582_1	NC_024254	join(88331..88532,88683..89524)
LOC103402582_2	NC_024254	join(88331..88532,88705..89524)
LOC103402581_1	NC_024254	join(95621..96140,96635..96876,96984..97106,97224..97340,97454..97636,97889..98030,98388..98626)
LOC103402586_1	NC_024254	join(101868..102096,102186..102642)
LOC103402312_1	NC_024254	join(106367..106480,107775..107993,108473..108583,108700..108768,108897..109040,109156..109230,109315..109491,109612..109697,109770..109857)
LOC103402587_1	NC_024254	join(111966..112404,112505..112599,112708..112855,113018..113109,113202..113301,113598..113664,113760..113880,114055..114171,114244..114301)
LOC103402588_1	NC_024254	complement(join(115403..115697,115791..115867,116125..116395,116611..116897,117320..117446,117549..117665,117766..117879,118480..118581))
LOC103402589_1	NC_024254	join(128752..129009,129115..129254,129363..129756)
LOC103402590_1	NC_024254	complement(join(132452..132531,132664..132758,132842..133075,133287..133371,133513..133680,133981..134140,134284..134339,134777..134804))
LOC103402590_2	NC_024254	complement(join(132452..132531,132664..132742,132842..133075,133287..133371,133513..133680,133981..134140,134284..134339,134777..134804))
LOC103402590_3	NC_024254	complement(join(132452..132531,132664..132742,132842..133075,133287..133371,133513..133680,133981..134140,134284..134339,134777..134804))
LOC103402590_4	NC_024254	complement(join(132452..132531,132664..132742,132842..133075,133287..133371,133513..133680,133981..134140,134284..134339,134777..134804))

結果

- Leaf1_ERR268472_1 (paired) trimmed (paired) (GE)
- Leaf1_ERR268472_1 (paired) trimmed (paired) (TE)
- Leaf1_ERR268472_1 (paired) trimmed (paired) RNA-Seq (Reads)
- Leaf1_ERR268472_1 (paired) trimmed (paired) RNA-Seq un-mapped reads [ERR268472_1] (single)
- Leaf1_ERR268472_1 (paired) trimmed (paired) RNA-Seq un-mapped reads [ERR268472_1] (paired)
- Leaf1_ERR268472_1 (paired) trimmed (paired) RNA-Seq report

マッピング トラック

遺伝子レベルでの発現量が表示されています。



結果

- Leaf1_ERR268472_1 (paired) trimmed (paired) (GE)
- Leaf1_ERR268472_1 (paired) trimmed (paired) (TE)
- Leaf1_ERR268472_1 (paired) trimmed (paired) RNA-Seq (Reads)
- Leaf1_ERR268472_1 (paired) trimmed (paired) RNA-Seq un-mapped reads [ERR268472_1] (single)
- Leaf1_ERR268472_1 (paired) trimmed (paired) RNA-Seq un-mapped reads [ERR268472_1] (paired)
- Leaf1_ERR268472_1 (paired) trimmed (paired) RNA-Seq report

レポート

Leaf1_ERR2684... X

1 Selected input sequences

1.1 Sequence reads

Name	Number of reads	Longest read	paired
Leaf1_ERR268472_1 (paired) trimmed (paired)	939,868	101	yes

For 'paired' data, there are two 'reads' in a pair.

1.2 Reference Sequences

References	Length	Genes	Transcripts
1	20,664,393	1,415	1,506

2 References

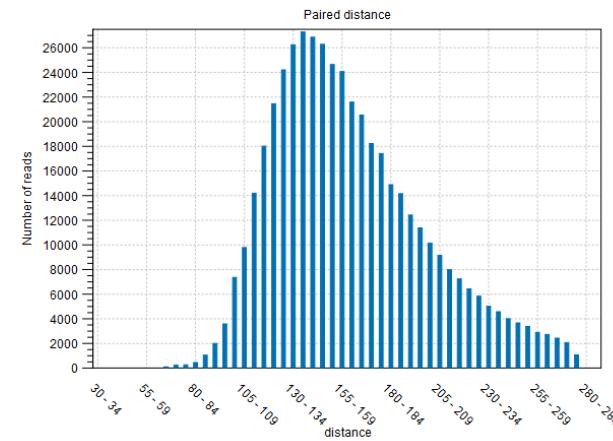
Input contained 1,415 genes and 1,506 transcripts.

2.1 Transcripts per gene

Transcripts per gene

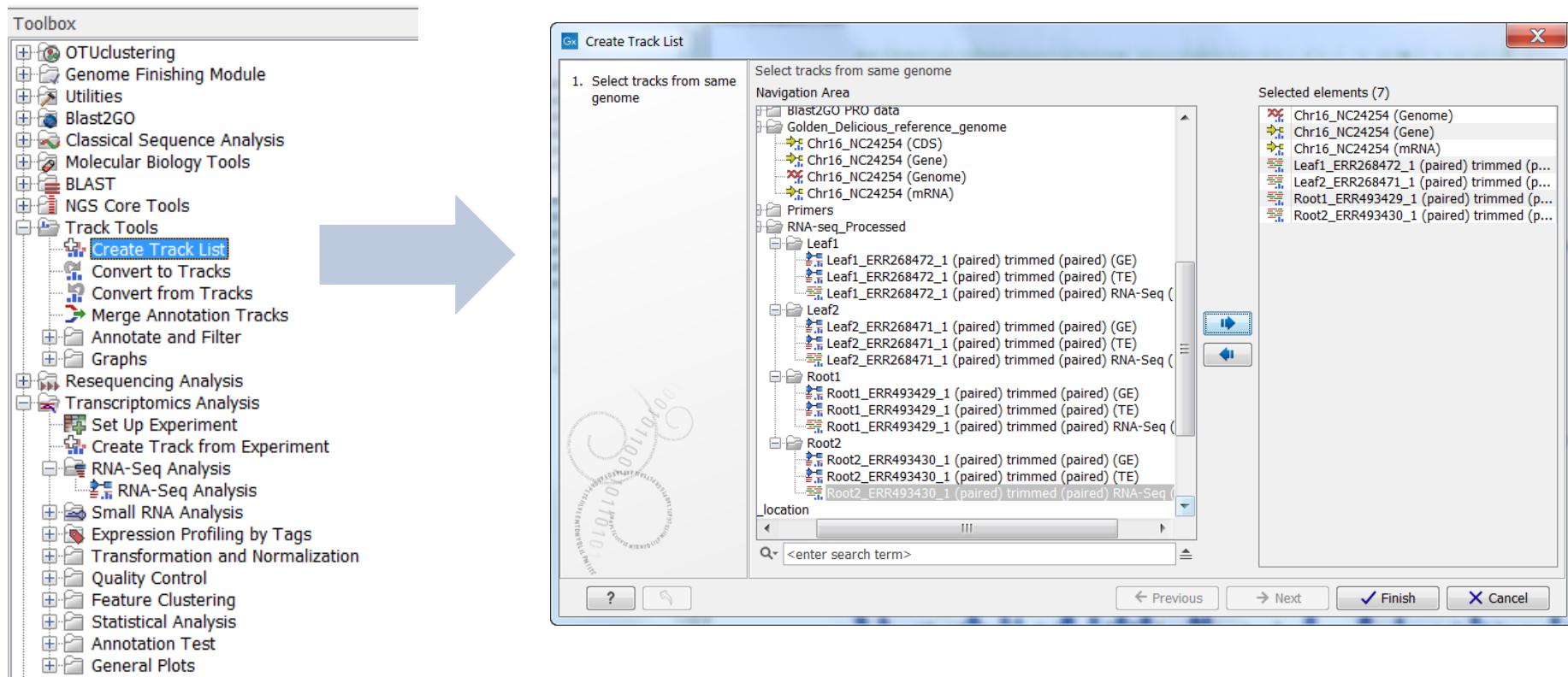
Transcripts	Number of genes
0	~150
1	~1000
2	~150
3	~20
4	~10
5	~5
6	~2
7	~1

3.5 Paired distance



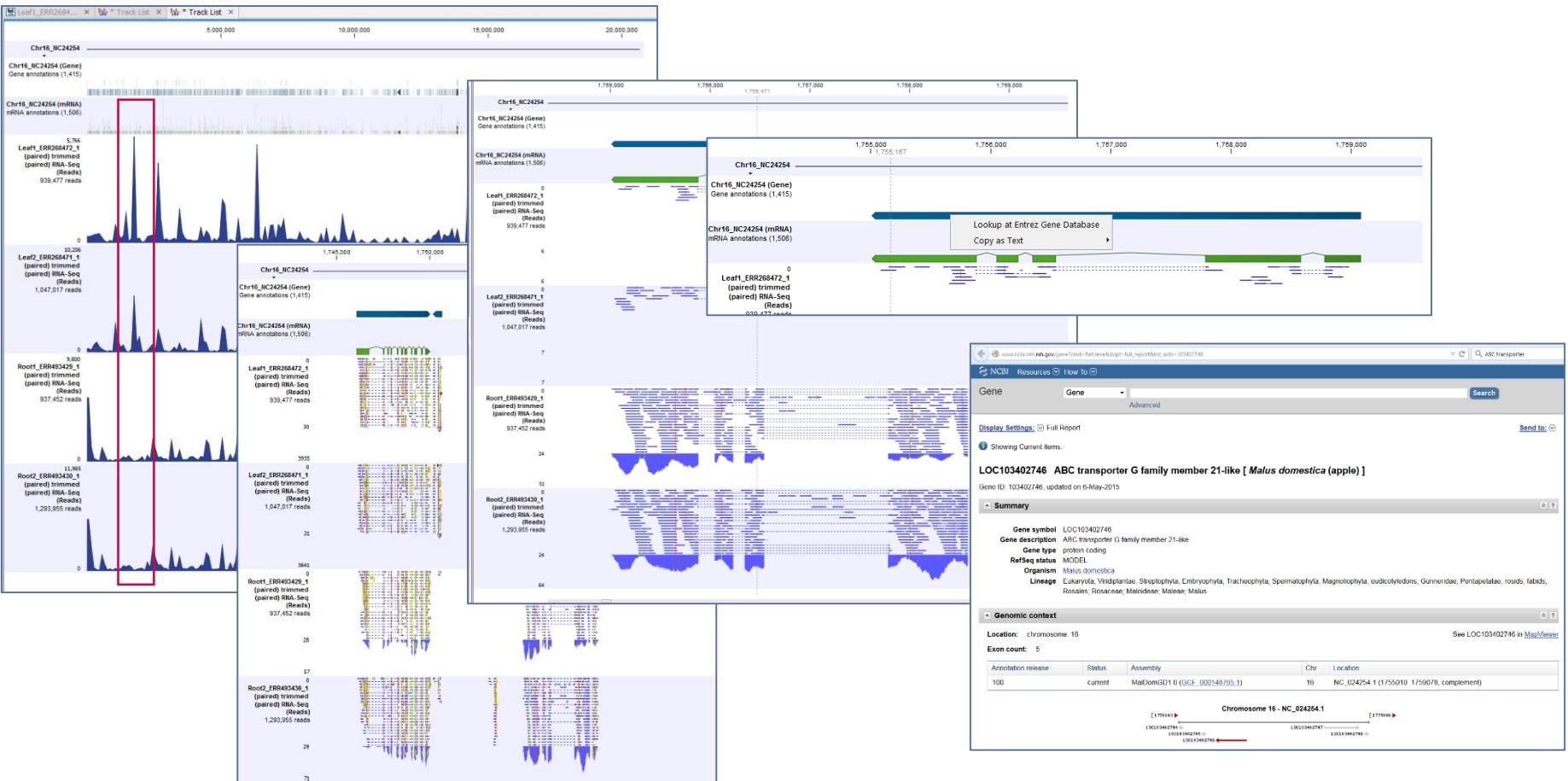
- 利用したデータのサマリーや、1つの遺伝子に対するトランスクriプトの数などの統計情報が含まれています。

Track list の作成



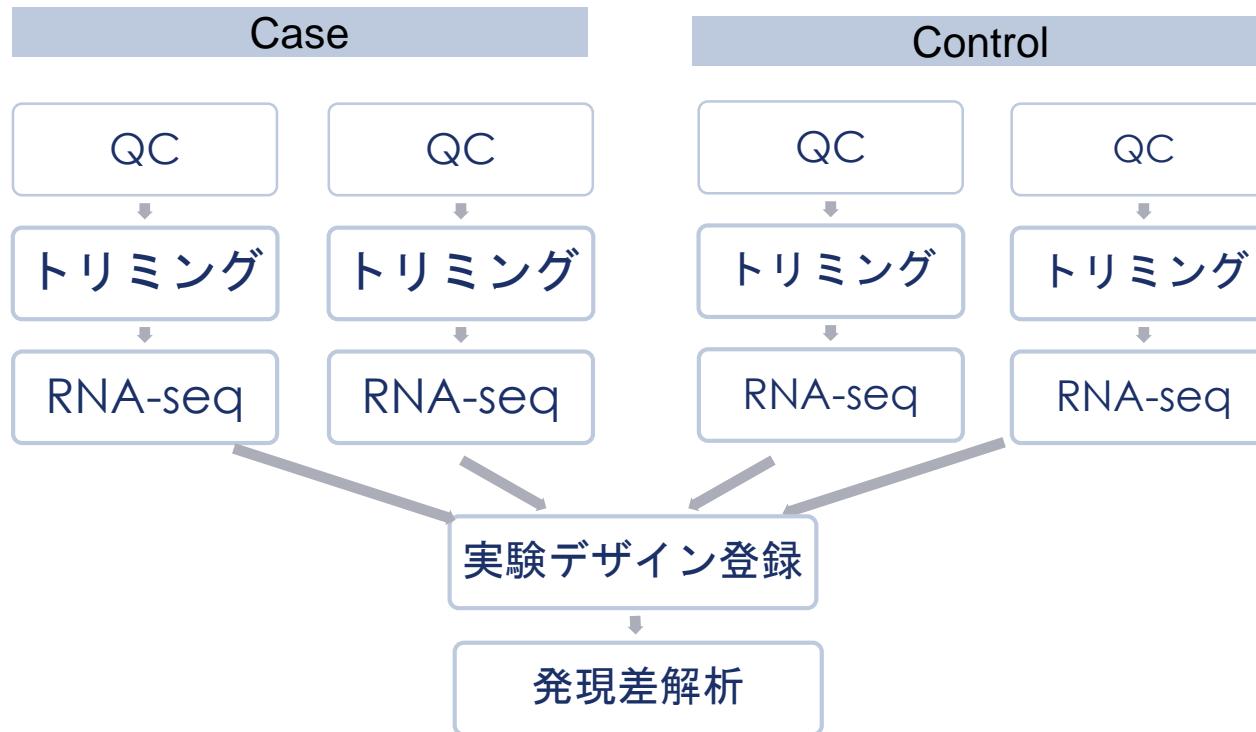
Toolbox > Track tools > Create track list を選択し、ゲノム、遺伝子、mRNA、マッピング（Reads）データを選択します。

Track list の作成



- ズームしていくと、詳細が確認できます。遺伝子のトラックから該当する遺伝子を右クリックし、Entrezでの検索も直接行えます。

複数処理

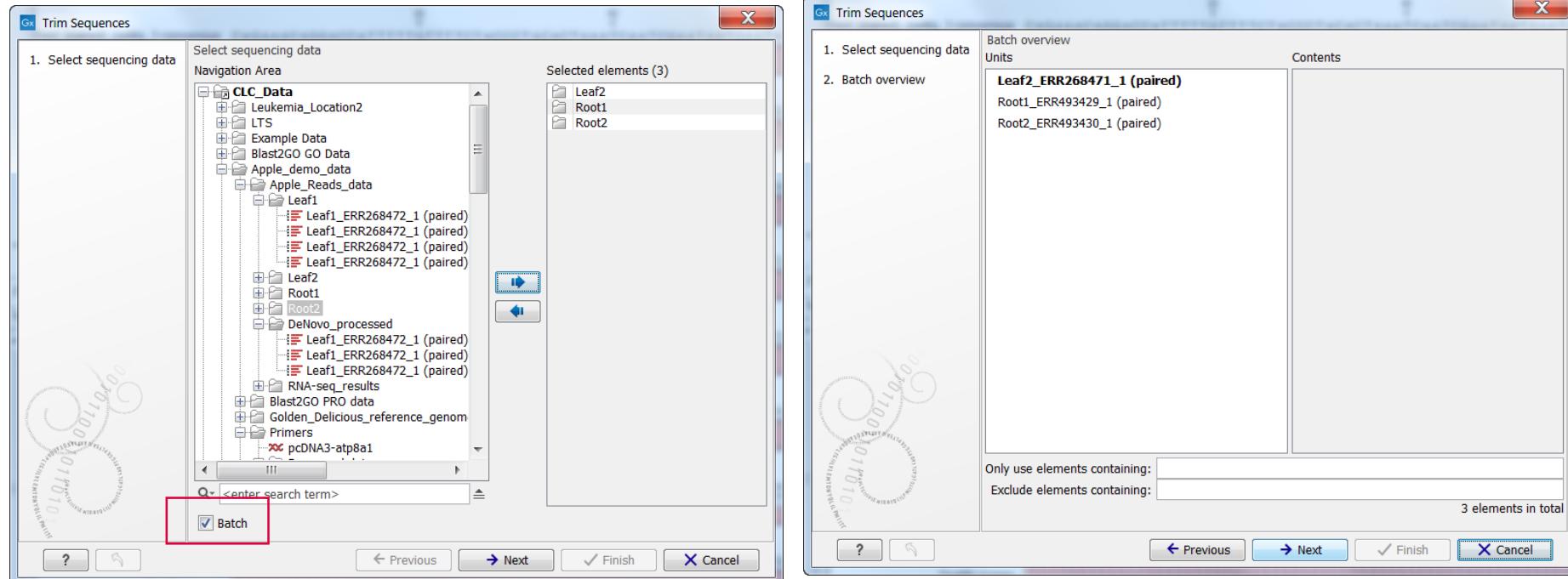


- 発現差解析では通常複数のサンプルに同じ処理が必要です。そのような場合に、バッチ処理とワークフローが活用できます。

CLC Genomics Workbench

バッチ処理

バッチ処理



- バッチ処理を用いることにより、複数のデータに対して同じ処理を一度に行うことができます
- ツールを立ち上げた際に、左下の Batch をチェックします
- バッチ処理するデータを含むフォルダーを選択します

```
RNA-seq_Processed
  └── Leaf1
      ├── Leaf1_ERR268472_1 (paired) trimmed (paired) (GE)
      ├── Leaf1_ERR268472_1 (paired) trimmed (paired) (TE)
      ├── Leaf1_ERR268472_1 (paired) trimmed (paired) RNA-Seq (Reads)
      ├── Leaf1_ERR268472_1 (paired) trimmed (paired) RNA-Seq report
      ├── Leaf1_ERR268472_1 (paired) trimmed (paired) RNA-Seq un-mapped reads [ERR268472_1] (paired)
      └── Leaf1_ERR268472_1 (paired) trimmed (paired) RNA-Seq un-mapped reads [ERR268472_1] (single)

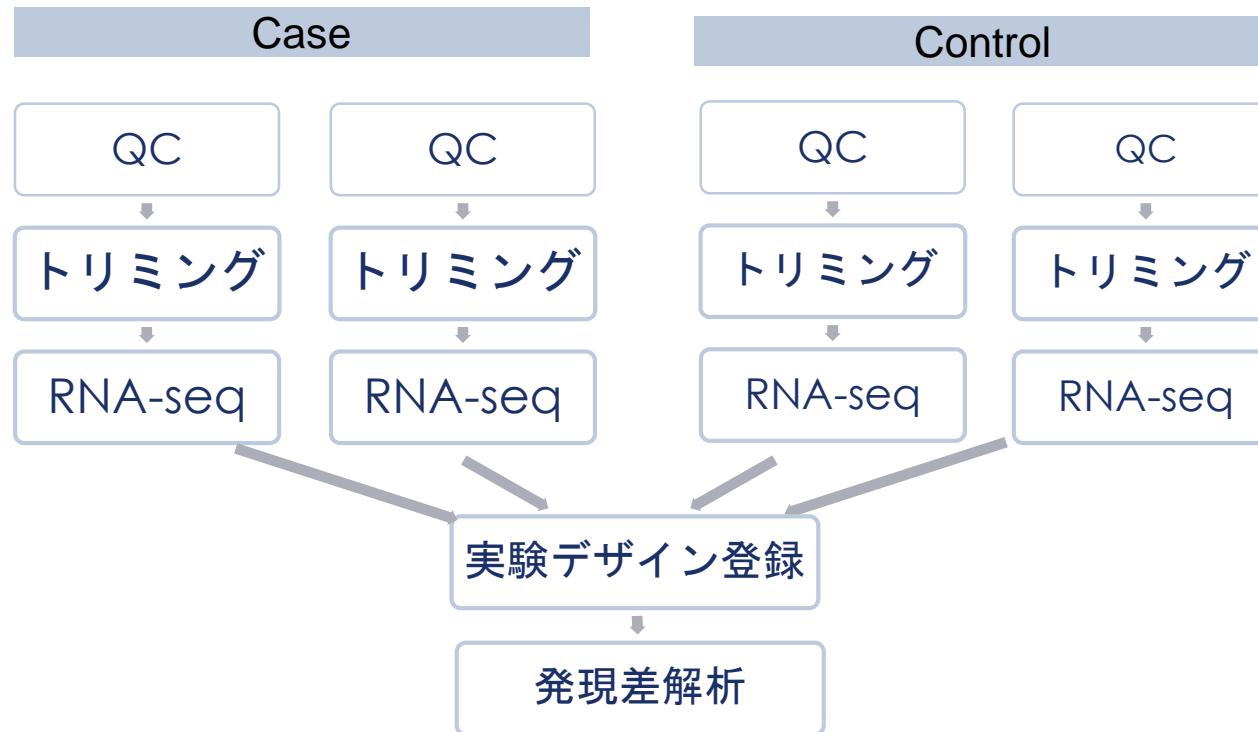
  └── Leaf2
      ├── Leaf2_ERR268471_1 (paired) trimmed (paired) (GE)
      ├── Leaf2_ERR268471_1 (paired) trimmed (paired) (TE)
      ├── Leaf2_ERR268471_1 (paired) trimmed (paired) RNA-Seq (Reads)
      ├── Leaf2_ERR268471_1 (paired) trimmed (paired) RNA-Seq report
      ├── Leaf2_ERR268471_1 (paired) trimmed (paired) RNA-Seq un-mapped reads [ERR268471_1] (paired)
      └── Leaf2_ERR268471_1 (paired) trimmed (paired) RNA-Seq un-mapped reads [ERR268471_1] (single)

  └── Root1
      ├── Root1_ERR493429_1 (paired) trimmed (paired) (GE)
      ├── Root1_ERR493429_1 (paired) trimmed (paired) (TE)
      ├── Root1_ERR493429_1 (paired) trimmed (paired) RNA-Seq (Reads)
      ├── Root1_ERR493429_1 (paired) trimmed (paired) RNA-Seq report
      ├── Root1_ERR493429_1 (paired) trimmed (paired) RNA-Seq un-mapped reads [ERR493429_1] (paired)
      └── Root1_ERR493429_1 (paired) trimmed (paired) RNA-Seq un-mapped reads [ERR493429_1] (single)

  └── Root2
      ├── Root2_ERR493430_1 (paired) trimmed (paired) (GE)
      ├── Root2_ERR493430_1 (paired) trimmed (paired) (TE)
      ├── Root2_ERR493430_1 (paired) trimmed (paired) RNA-Seq (Reads)
      ├── Root2_ERR493430_1 (paired) trimmed (paired) RNA-Seq report
      ├── Root2_ERR493430_1 (paired) trimmed (paired) RNA-Seq un-mapped reads [ERR493430_1] (paired)
      └── Root2_ERR493430_1 (paired) trimmed (paired) RNA-Seq un-mapped reads [ERR493430_1] (single)
```

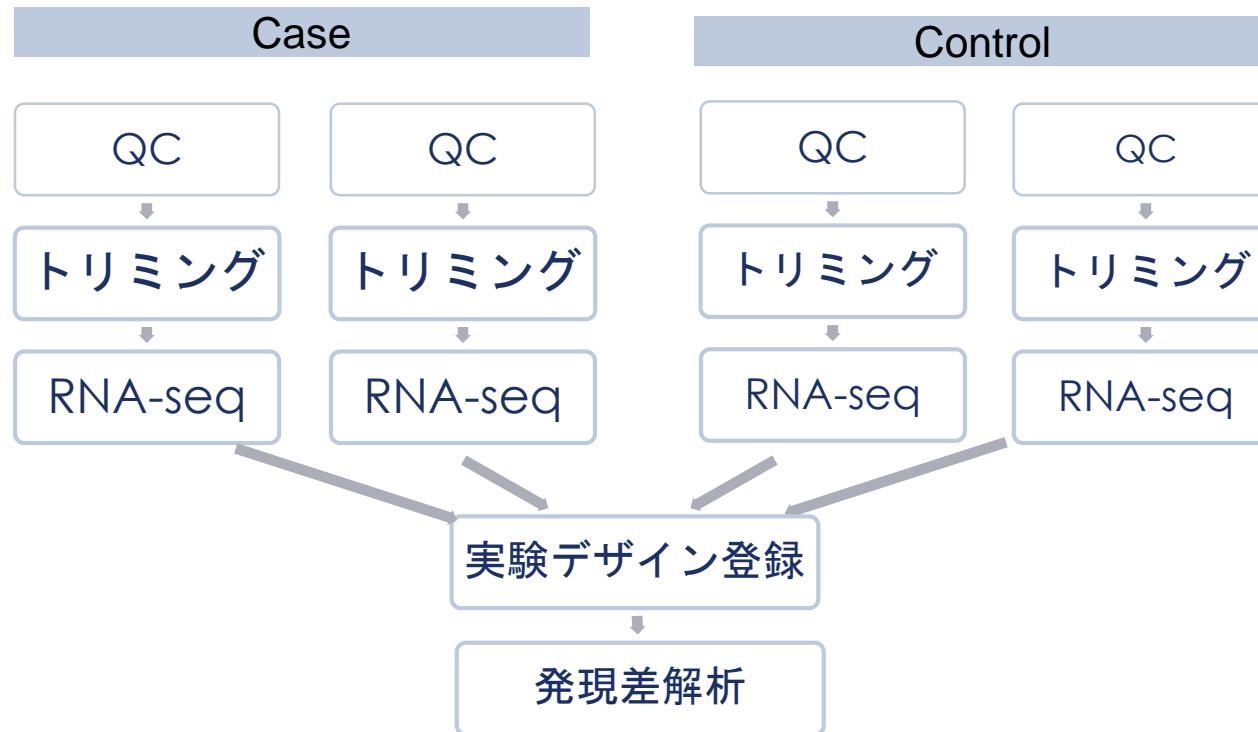
- ツールのパラメータを設定して実行すると、選択した全てのデータに対して一度に処理が行われます
- データがそれぞれ別のフォルダーにある場合には、それぞれのフォルダーに分けて結果が出力されます
- 次に説明する workflow と組み合わせることで、一度に複数のサンプルを同じワークフローで解析することができます

ワークフローツール



発現差解析

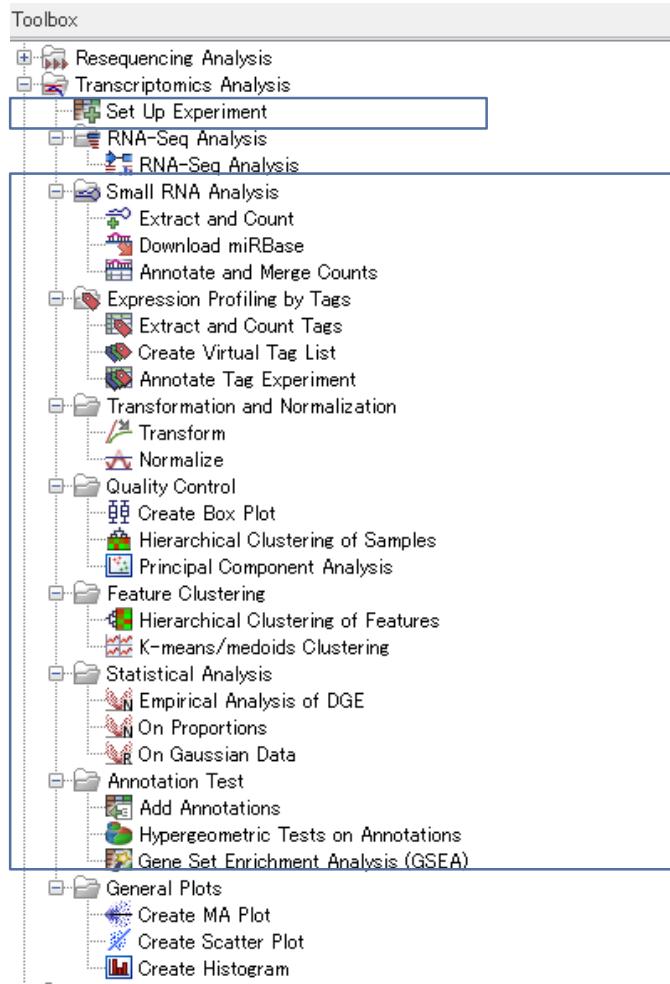
実験デザイン



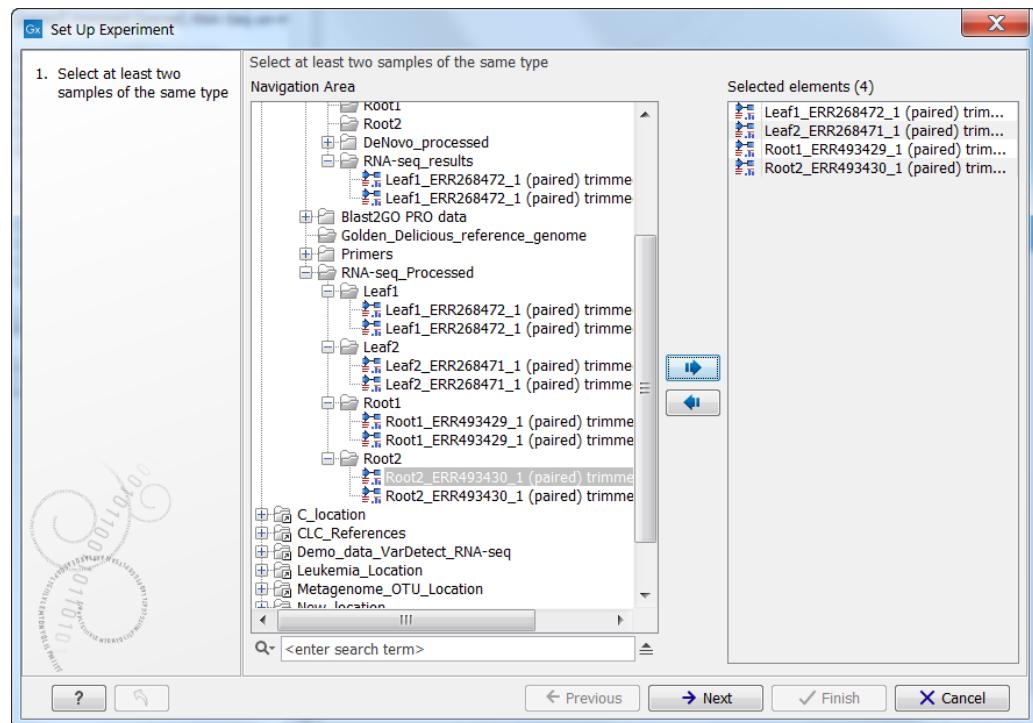
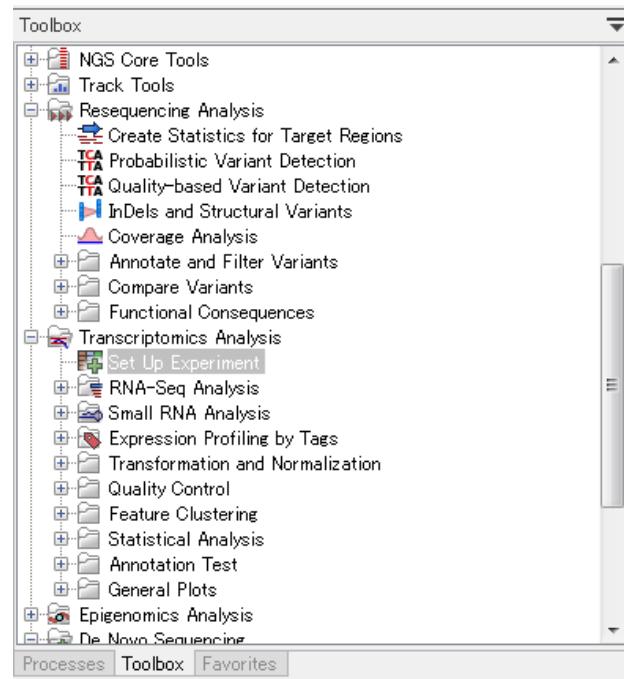
- 実験デザインをGenomics Workbenchへ登録し統計検定により、発現差のある遺伝子を見つけ出す。

Expression analysis – Set Up Experiment

- 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。

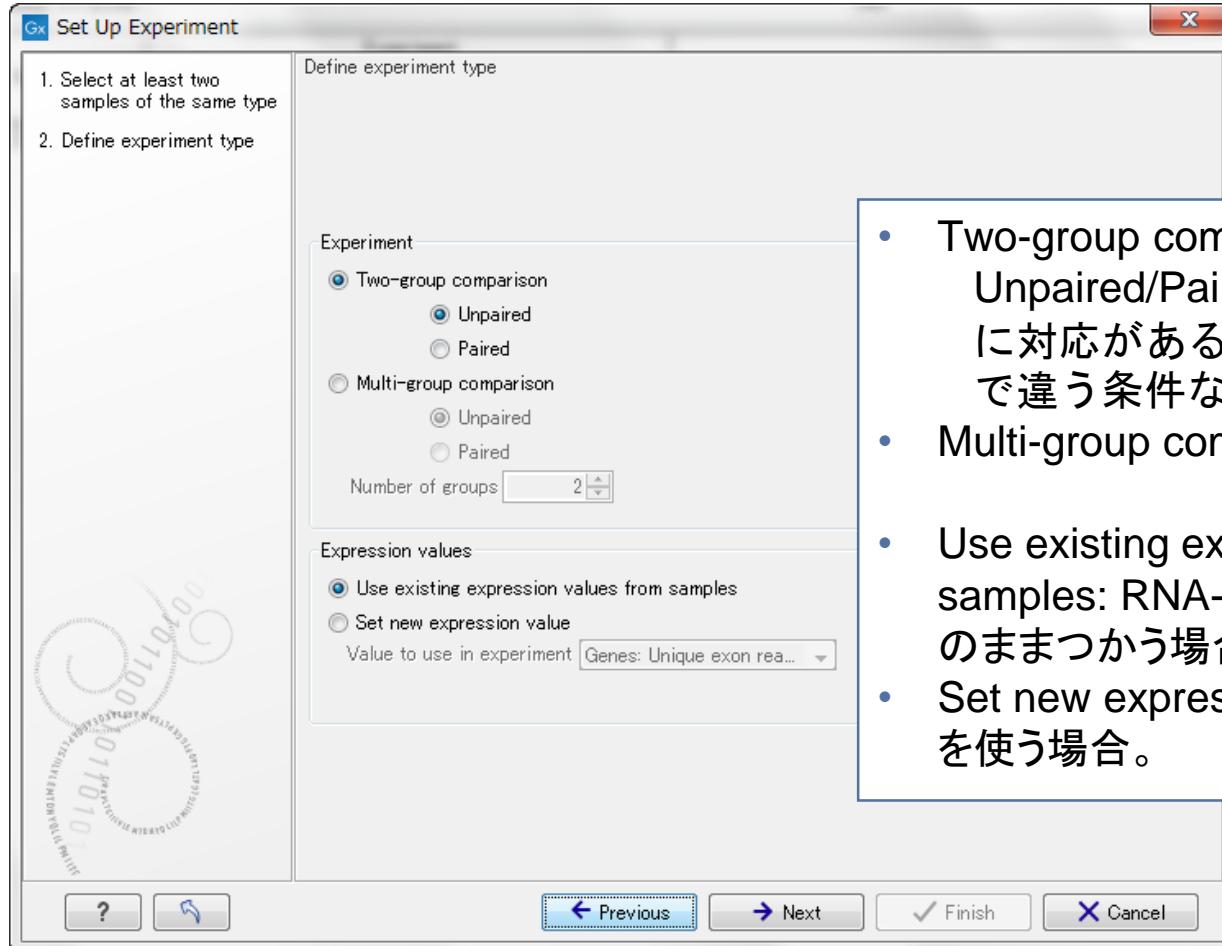


Set Up Experiment



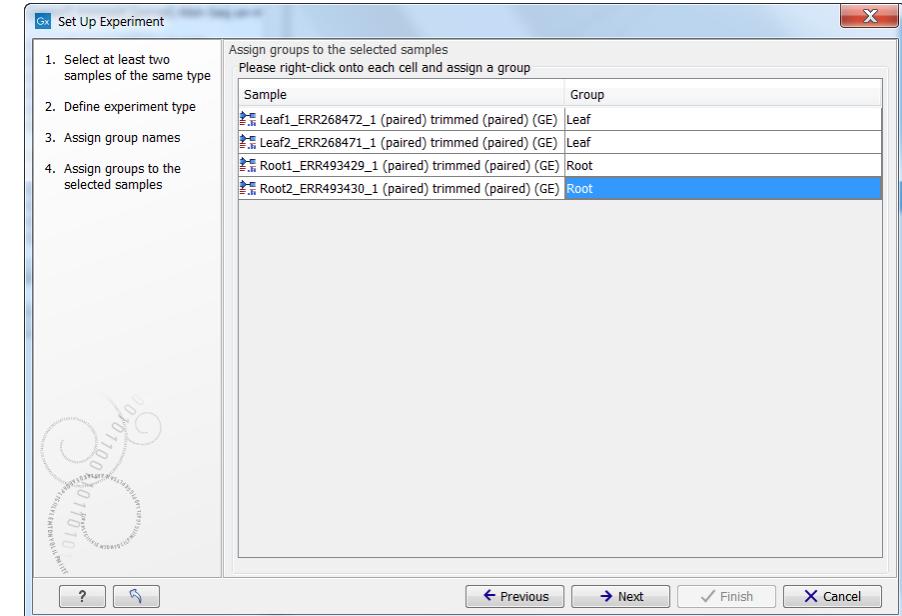
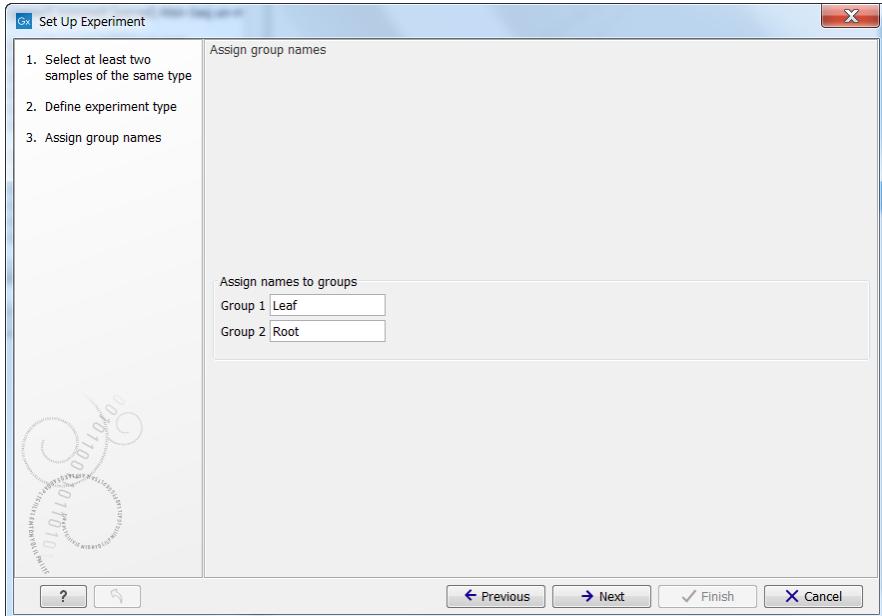
- Toolboxから Set Up Experiment を選択、ダブルクリック。
- RNA-seqの結果から、GEのトラック4種類を選択。

Set Up Experiment



- **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。

Set Up Experiment



- 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

Set Up Experiment

Expression analysis – Set Up Experiment

Leaf vs. Root

Rows: 1,415

Feature ID	Experiment				Analysis					
	Range (original values)	IQR (original values)	Difference (...)	Fold Chang...	Expression values	Transcript...	Detected t...	Exon length	Unique ge...	Total
LOC103402574	3.00	2.00	-1.00	-2.00	1.00	1	1	759	1	
LOC103402575	208.00	208.00	-177.50	-∞	147.00	1	1	533	151	
LOC103402576	2.00	2.00	1.00	∞	0.00	1	0	1000	0	
LOC103402577	0.00	0.00	0.00	1.00	0.00	1	0	774	0	
LOC103402578	0.00	0.00	0.00	1.00	0.00	0	0	0	9	
LOC103402579	809.00	808.00	-768.50	-1,538.00	729.00	1	1	809	731	
LOC103402580	247.00	220.00	215.50	3.26	82.00	1	1	2490	88	
LOC103402581	123.00	114.00	-31.50	-1.23	105.00	1	1	1566	106	
LOC103402582	5.00	3.00	4.00	5.00	0.00	2	0	1044	1	
LOC103402583	490.00	466.00	-439.50	-7.19	472.00	3	0	1492	492	
LOC103402584	146.00	105.00	-17.50	-1.04	534.00	1	1	1361	539	
LOC103402586	5,660.00	5,644.00	4,179.50	465.39	17.00	1	1	686	18	
LOC103402587	531.00	326.00	311.50	2.02	202.00	1	1	1614	207	
LOC103402588	2,312.00	2,296.00	-1,867.50	-170.77	2,315.00	1	1	2985	2327	
LOC103402589	2,630.00	2,481.00	2,147.50	14.30	87.00	1	1	792	88	
LOC103402590	111.00	11.00	-54.00	-1.37	207.00	5	2	1774	212	
LOC103402591	113.00	80.00	-30.50	-1.38	71.00	1	1	905	73	
LOC103402592	54.00	50.00	41.50	7.92	8.00	1	1	815	8	
LOC103402594	499.00	435.00	31.50	1.06	649.00	1	1	1572	560	
LOC103402595	0.00	0.00	0.00	1.00	0.00	0	0	0	89	
LOC103402596	456.00	411.00	-246.00	-4.81	123.00	1	1	3772	129	
LOC103402597	7.00	4.00	3.50	2.40	1.00	1	1	962	2	
LOC103402598	160.00	148.00	-129.00	-4.58	140.00	2	2	2637	144	
LOC103402599	0.00	0.00	0.00	1.00	0.00	0	0	0	182	
LOC103402600	235.00	179.00	-138.50	-2.05	202.00	1	1	671	202	

Experiment Table Settings

- IQR (original values)
- Difference (original values)
- Fold Change (original values)
-
-

Analysis level

-
-

Annotation level

Group level

- Leaf
- Root
- Group columns
- Means
-
-

Sample level

- 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 reads
- Total intron reads
- Ratio of intron to total gene reads
- Exons
- RPKM
- Chromosome
- Chromosome region start
- Chromosome region end
-
-

- 結果には、各サンプルのそれぞれの統計値やグループ内での統計値、合計した統計値が含まれる。

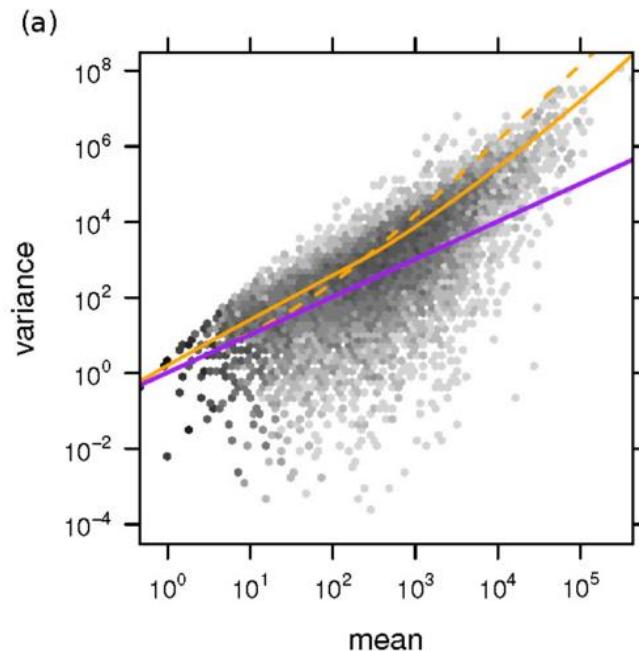
Expression analysis

- RNA-seqの結果を使から、「RootとLeafでの違いを調べる」という事を行います。
- 7.0から新しく搭載されたEmpirical Analysis of DGEについて使い方を説明します。

		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
	Bagglaley's test	2群	必須
Empirical Analysis of DGE		2群	必須

Expression analysis -EDGE

- カウントデータで用いられる統計分布は、ポアソン分布が多くありました（理由としては、平均値=分散という分布のため、推定するパラメータが1つでよいため）、多くのデータで、分散が平均よりも大きくなることが観察され、より適した分布として負の二項分布が利用されるようになりました。



Anders S, Huber W: Differential expression analysis for sequence count data. *Genome Biol* 2010, 11:R106.

Expression analysis -EDGE

- カウントデータを使い、データが負の二項分布に従うと仮定して必要なパラメータ（平均値とDispersion）を推定し、Exact Testで検定を行います。
- FisherのExact Testはよく論文などで利用され、分割表から考えることができます。

	条件1(例 : Leaf)	条件2 (例 : Root)	合計
遺伝子 A	n_{11}	n_{12}	$n_{11} + n_{12}$
残りの遺伝子	n_{21}	n_{22}	$n_{21} + n_{22}$
合計	$n_{11} + n_{21}$	$n_{12} + n_{22}$	N

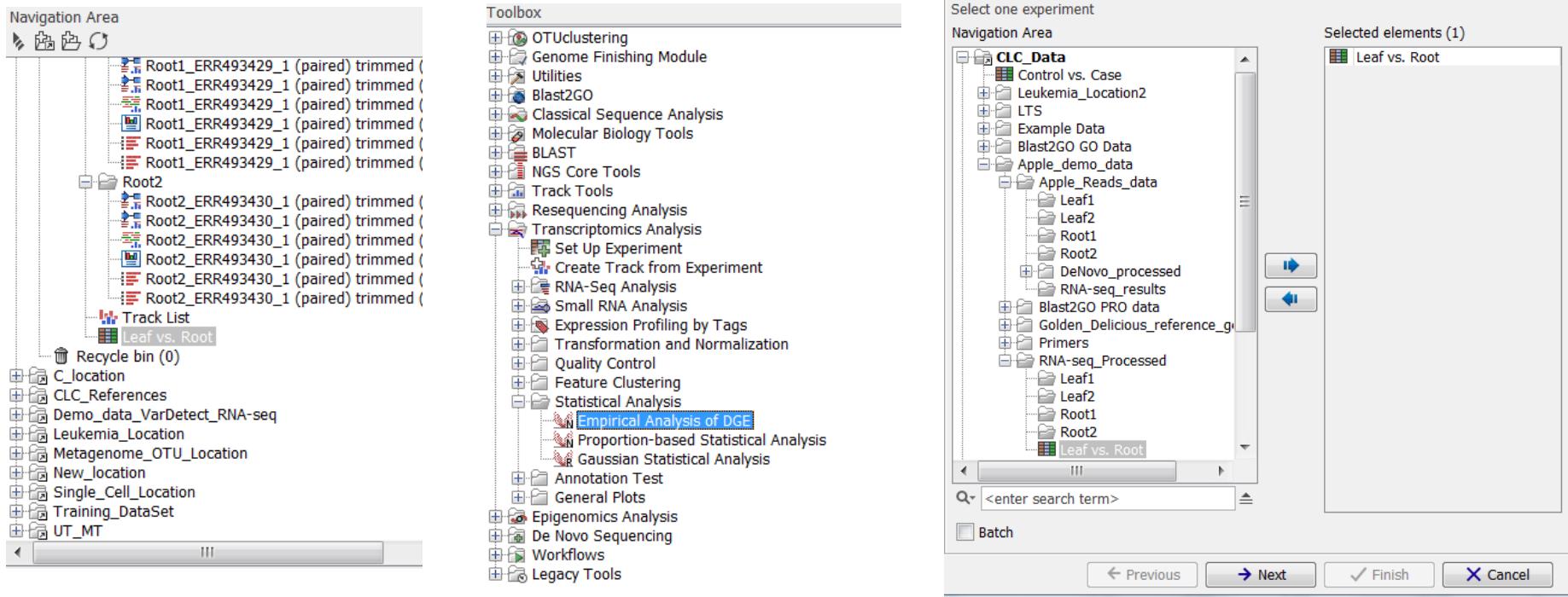
$$p = \frac{n_{11}+n_{12} C_{n_{11}} \cdot n_{21}+n_{22} C_{n_{21}}}{N C_{n_{11}+n_{12}}}$$

- FisherのExact Testは超幾何分布を仮定していますが、EdgeRでは負の二項分布を使い、レプリケートに対応しています。
- Rに搭載されているEdgeRでは、Exact Testと複数の要因の検定に対応できるようGLM(Generalized Linear Model)が搭載されていますが、Genomics Workbenchでは、Exact Testのみが現在は搭載されています。

その他の方法

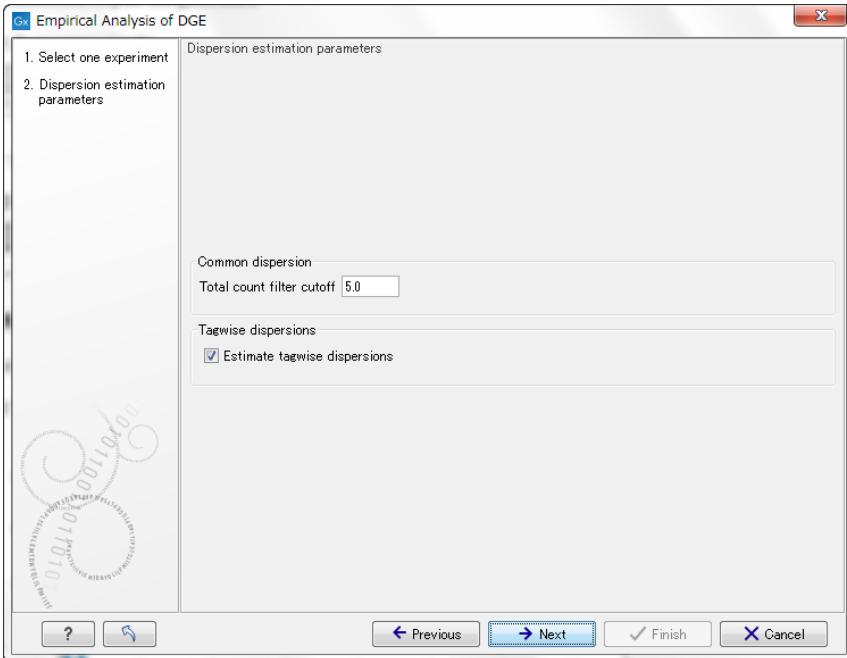
手法	仮定する分布	ノーマライズ法	検定
edgeR	負の二項分布	TMM/Upper quantile/RLE	Exact Test/一般化線形モデル
DESeq2	負の二項分布	Size Factor	一般化線形モデル
baySeq	負の二項分布	Size Factor	ベイズ法
Cuffdiff 2 (Cufflinks)	ベータ負の二項分布	Geometric/Upper quantile/FPKM	t検定

Expression analysis - EDGE

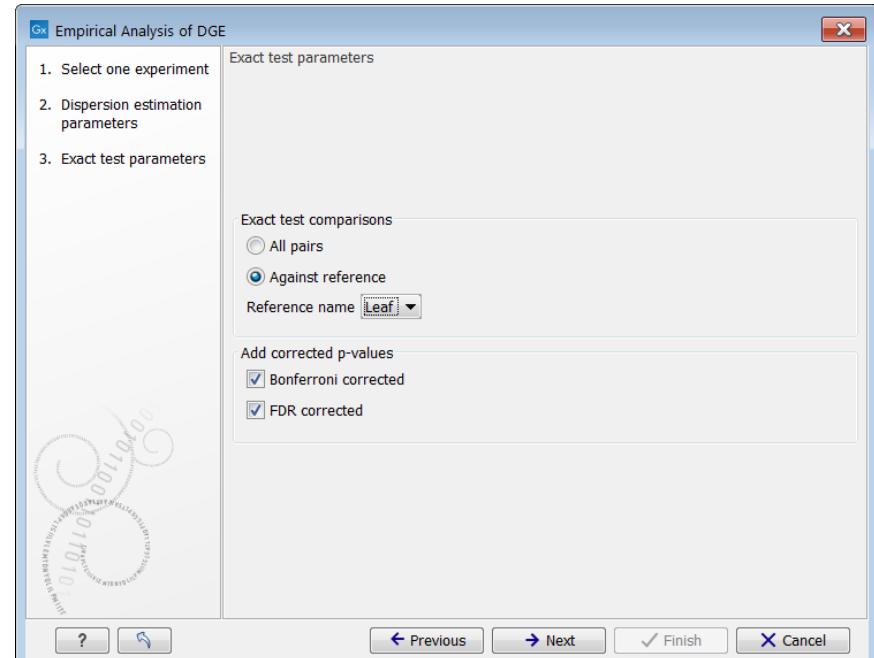


- Navigation Areaから使用するExperimentデータを選択。
- Toolboxから Empirical Analysis of DGE を選択、ダブルクリック。
- ウィザードが起動し、選択したデータが選ばれていることを確認。

Expression analysis - EDGE



- Total count filter cutoff: 発現量があるとするための最小のカウント数
- Tagwise dispersions: タグごと(Set Up Experimentを遺伝子レベルで作成した場合は、遺伝子ごと)のばらつきを計算させるか。通常はチェックをいれたままにしておいてください。



- Exact test comparisons
すべての組み合わせ
指定したものをコントロールとする場合
- Add corrected p-value: p値の補正
ボンフェローニ
FDR

Expression analysis - EDGE

Rows: 1,415										Filter
Feature ID	Experiment				EDGE test: Leaf vs Root, tagwise dispersions					Exp
	Range (original values)	IQR (original values)	Difference (...)	Fold Chang...	P-value	Fold change	Weighted difference	Bonferroni	FDR p-value correction	
LOC103402574	3.00	2.00	-1.00	-2.00	0.46	-2.48	-2.85E-6	1.00	0.76	
LOC103402575	208.00	208.00	-177.50	-0.00	1.25E-16	-1,625.58	-4.03E-4	1.77E-13	5.19E-15	
LOC103402576	2.00	2.00	1.00	0.00	0.55	8.65	1.90E-6	1.00	0.85	
LOC103402577	0.00	0.00	0.00	1.00	1.00	-1.00	-4.24E-22	1.00	1.00	
LOC103402578	0.00	0.00	0.00	1.00	1.00	-1.00	-4.24E-22	1.00	1.00	
LOC103402579	809.00	808.00	-768.50	-1,538.00	3.15E-30	-1,571.11	-1.75E-3	4.46E-27	4.96E-28	
LOC103402580	247.00	220.00	215.50	3.25	0.01	2.58	3.43E-4	1.00	0.04	
LOC103402581	123.00	114.00	-31.50	-1.23	0.27	-1.57	-1.37E-4	1.00	0.51	
LOC103402582	5.00	3.00	4.00	5.00	0.17	3.68	6.72E-6	1.00	0.36	
LOC103402583	490.00	466.00	-439.50	-7.19	3.97E-9	-9.27	-1.04E-3	5.61E-6	5.85E-8	
LOC103402584	146.00	105.00	-17.50	-1.04	0.41	-1.34	-2.85E-4	1.00	0.70	
LOC103402586	5,660.00	5,644.00	4,179.50	465.30	1.33E-16	335.31	7.10E-3	1.88E-13	5.38E-15	
LOC103402587	531.00	326.00	311.50	2.02	0.22	1.57	3.95E-4	1.00	0.43	
LOC103402588	2,312.00	2,296.00	-1,867.50	-170.77	2.32E-21	-236.70	-4.31E-3	3.28E-18	1.49E-19	
LOC103402589	2,630.00	2,481.00	2,147.50	14.30	4.56E-8	11.19	3.70E-3	6.45E-5	5.52E-7	
LOC103402590	111.00	11.00	-54.00	-1.37	0.11	-1.83	-2.09E-4	1.00	0.26	
LOC103402591	113.00	80.00	-30.50	-1.33	0.18	-1.89	-1.18E-4	1.00	0.37	
LOC103402592	54.00	50.00	41.50	7.92	1.82E-3	5.93	6.92E-5	1.00	7.91E-3	
LOC103402594	499.00	435.00	31.50	1.06	0.50	-1.31	-2.69E-4	1.00	0.81	
LOC103402595	0.00	0.00	0.00	1.00	1.00	-1.00	-4.24E-22	1.00	1.00	
LOC103402596	456.00	411.00	-246.00	-4.81	1.50E-4	-6.30	-5.84E-4	0.21	8.54E-4	
LOC103402597	7.00	4.00	3.50	2.40	0.61	1.80	4.70E-6	1.00	0.93	
LOC103402598	160.00	148.00	-129.00	-4.53	4.85E-5	-5.89	-3.11E-4	0.07	3.12E-4	
LOC103402599	0.00	0.00	0.00	1.00	1.00	-1.00	-4.24E-22	1.00	1.00	
LOC103402600	235.00	179.00	-138.50	-2.05	8.82E-3	-2.65	-3.81E-4	1.00	0.03	
LOC103402601	145.00	99.00	16.00	1.02	0.61	-1.21	-2.80E-4	1.00	0.92	
LOC103402602	9.00	7.00	-4.50	-5.50	0.05	-6.77	-1.11E-5	1.00	0.13	
LOC103402603	243.00	215.00	180.50	2.09	0.18	1.62	2.34E-4	1.00	0.38	
LOC103402604	15.00	14.00	9.50	20.00	8.88E-3	12.31	1.56E-5	1.00	0.03	
LOC103402605	0.00	0.00	0.00	1.00	1.00	-1.00	-4.24E-22	1.00	1.00	
LOC103402606	288.00	285.00	133.00	1.47	0.81	1.09	6.11E-5	1.00	1.00	
LOC103402607	274.00	249.00	107.50	1.30	0.98	-1.01	-8.91E-6	1.00	1.00	
LOC103402608	1,220.00	1,183.00	1,150.50	20.02	7.23E-11	15.97	2.05E-3	1.02E-7	1.40E-9	
LOC103402609	2,253.00	2,201.00	-1,602.50	-17.87	4.69E-10	-21.06	-3.64E-3	6.64E-7	7.90E-9	
LOC103402610	311.00	302.00	258.00	28.16	1.12E-9	21.69	4.49E-4	1.58E-6	1.82E-8	

結果

- 黒い▼ボタンをクリック

Rows: 1,415

Feature ID	Experiment				EDGE test: Leaf vs Root, tagwise dispersions					Ex
	Range (original values)	IQR (original values)	Difference (...)	Fold Chang...	P-value	Fold change	Weighted difference	Bonferroni	FDR p-value correction	
LOC103402574	3.00	2.00	-1.00	-2.00	0.46	-2.48	-2.85E-6	1.00	0.76	
LOC103402575	208.00	208.00	-177.50	-∞	1.25E-16	-1,625.58	-4.03E-4	1.77E-13	5.19E-15	
LOC103402576	2.00	2.00	1.00	∞	0.55	8.65	1.90E-6	1.00	0.85	
LOC103402577	0.00	0.00	0.00	1.00	1.00	-1.00	-4.24E-22	1.00	1.00	
LOC103402578	0.00	0.00	0.00	1.00	1.00	-1.00	-4.24E-22	1.00	1.00	
LOC103402579	809.00	808.00	-768.50	-1,538.00	3.15E-30	-1,571.11	-1.75E-3	4.46E-27	4.96E-28	
LOC103402580	247.00	220.00	215.50	3.26	0.01	2.58	3.43E-4	1.00	0.04	
LOC103402581	123.00	114.00	-31.50	-1.23	0.27	-1.57	-1.37E-4	1.00	0.51	
LOC103402582	5.00	3.00	4.00	5.00	0.17	3.68	6.72E-6	1.00	0.36	
LOC103402583	490.00	466.00	-439.50	-7.19	3.97E-9	-9.27	-1.04E-3	5.61E-6	5.85E-8	
LOC103402584	146.00	105.00	-17.50	-1.04	0.41	-1.34	-2.85E-4	1.00	0.70	

Rows: 438 / 1,415

EDGE test: Leaf vs Root, tagwise dispersions - FDR p-value correction

Match any Match all

0.05

Filter

Feature ID	Experiment				FDR test: Leaf vs Root, tagwise dispersions					Ex
	Range (original values)	IQR (original values)	Difference (...)	Fold Chang...	P-value	Fold change	Weighted difference	Bonferroni	FDR p-value correction	
LOC103402575	208.00	208.00	-177.50	-∞	1.25E-16	-1,625.58	-4.03E-4	1.77E-13	5.19E-15	
LOC103402579	809.00	808.00	-768.50	-1,538.00	3.15E-30	-1,571.11	-1.75E-3	4.46E-27	4.96E-28	
LOC103402580	247.00	220.00	215.50	3.26	0.01	2.58	3.43E-4	1.00	0.04	
LOC103402583	490.00	466.00	-439.50	-7.19	3.97E-9	-9.27	-1.04E-3	5.61E-6	5.85E-8	
LOC103402586	5,660.00	5,644.00	4,179.50	465.39	1.33E-16	335.31	7.10E-3	1.88E-13	5.38E-15	
LOC103402588	2,312.00	2,296.00	-1,867.50	-170.77	2.32E-21	-236.70	-4.31E-3	3.28E-18	1.49E-19	
LOC103402589	2,630.00	2,481.00	2,147.50	14.30	4.56E-8	11.19	3.70E-3	6.45E-5	5.52E-7	
LOC103402592	54.00	50.00	41.50	7.92	1.82E-3	5.93	6.92E-5	1.00	7.91E-3	
LOC103402596	456.00	411.00	-424.00	-4.81	1.50E-4	-6.30	-5.84E-4	0.21	8.54E-4	
LOC103402598							1E-4	0.07	3.12E-4	
LOC103402600							1E-4	1.00	0.03	
LOC103402604							6E-5	1.00	0.03	
LOC103402608	1,220.00	1,183.00	1,150.50	20.02	7.23E-11	15.97	2.05E-3	1.02E-7	1.40E-9	
LOC103402609	2,253.00	2,201.00	-1,602.50	-17.87	4.69E-10	-21.06	-3.64E-3	6.64E-7	7.90E-9	
LOC103402610	311.00	302.00	258.00	28.16	1.12E-9	21.69	4.49E-4	1.58E-6	1.82E-8	
LOC103402611	517.00	513.00	-434.50	145.83	4.37E-18	107.76	7.60E-4	6.10E-15	2.38E-16	

- 列ごとのフィルター用のツールが現れます。



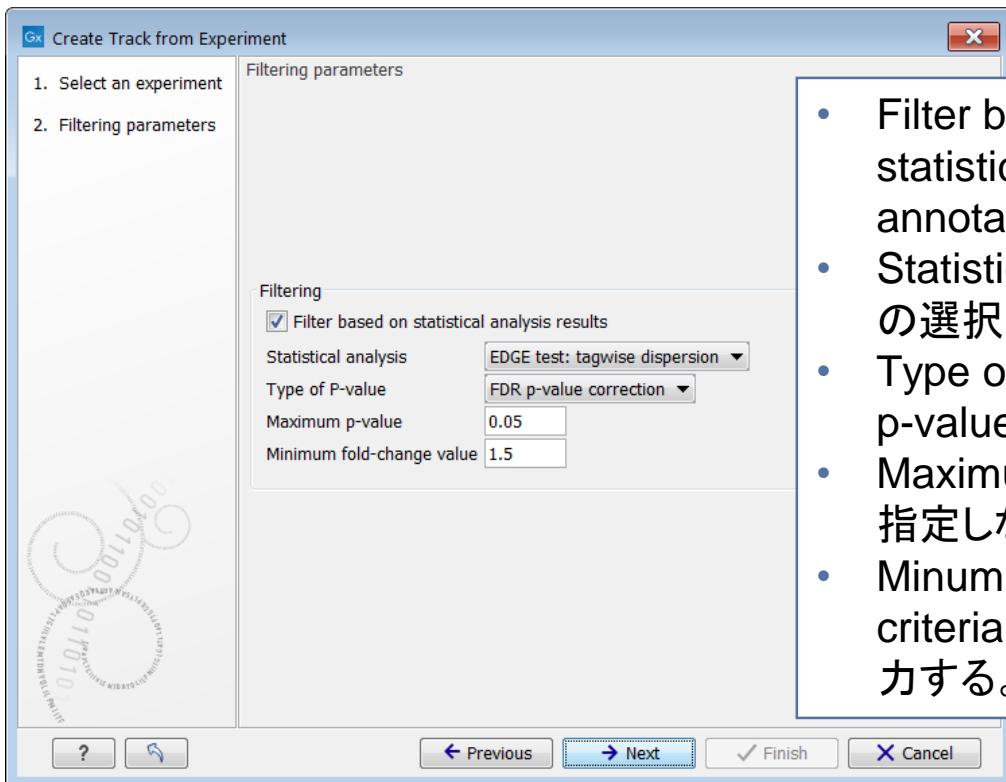
RNA-seq: 解析結果の Experiment から track の作成

The screenshot displays the CLC BioWorkshop software interface with three main panels:

- Navigation Area** (Left): Shows a tree view of project files and references. Projects include "Root1_ERR493429_1", "Root2", "Apple_demo_data-1", and "Recycle bin (0)". References like "C_location", "CLC_References", and "Training_DataSet" are also listed.
- Toolbox** (Center): A list of analysis tools categorized under "Transcriptomics Analysis". Tools include "Set Up Experiment", "Create Track from Experiment", "RNA-Seq Analysis", "Small RNA Analysis", "Expression Profiling by Tags", "Transformation and Normalization", "Quality Control", "Feature Clustering", "Statistical Analysis" (with sub-options for Empirical Analysis of DGE, Proportion-based Statistical Analysis, and Gaussian Statistical Analysis), "Annotation Test", "General Plots", "Epigenomics Analysis", "De Novo Sequencing", "Workflows", and "Legacy Tools".
- Select an experiment** (Right): A tree view of experimental data. The "Apple_demo_data-1" project is expanded, showing "Apple_Reads_data" which contains "Leaf1", "Leaf2", "Root1", and "Root2". "Root1" is further expanded to show "DeNovo_processed" and "RNA-seq_results". "Leaf vs. Root" is highlighted in the tree. Navigation buttons "Previous", "Next", "Finish", and "Cancel" are at the bottom.

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

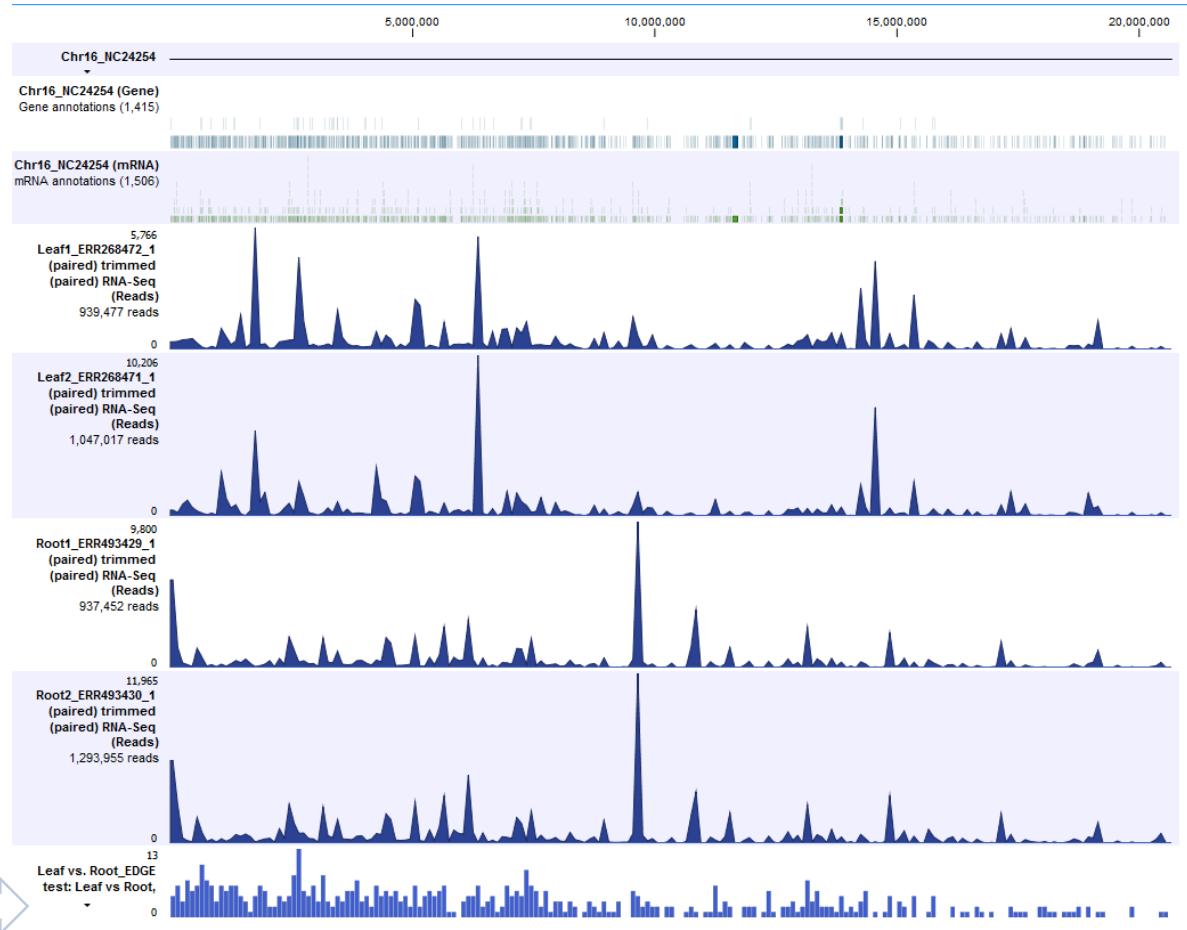
解析結果の Experiment から track の作成

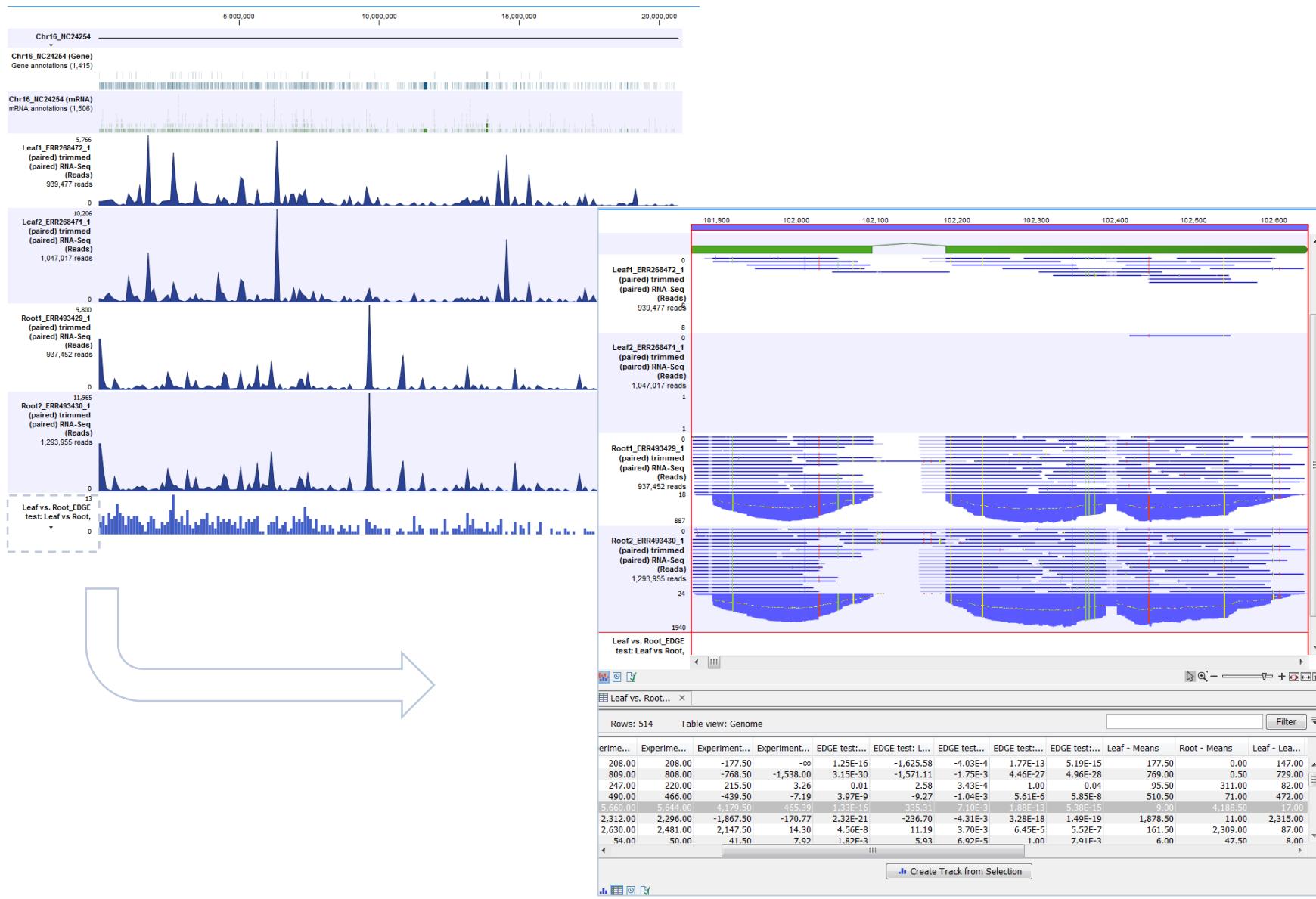


- Filter based on statiscal analysis results: statistical analysisで指定した基準を満たす annotationのみを track に移行する
- Statistical analysis: 使用する statistical analysis の選択
- Type of P-valew: raw p-value または corrected p-value のなかで使用するものを選択する
- Maximum p-value: p-palue の criteriaを指定する。指定しない場合には、1と入力する。
- Minumum fold-change value: fold change の criteria を指定する。指定しない場合には、0と入力する。

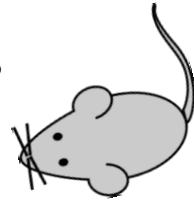
解析結果の Experiment から track の作成

Leaf vs. Root_EDGE test: Leaf vs Root, tagwise dispersions





お疲れ様でした。

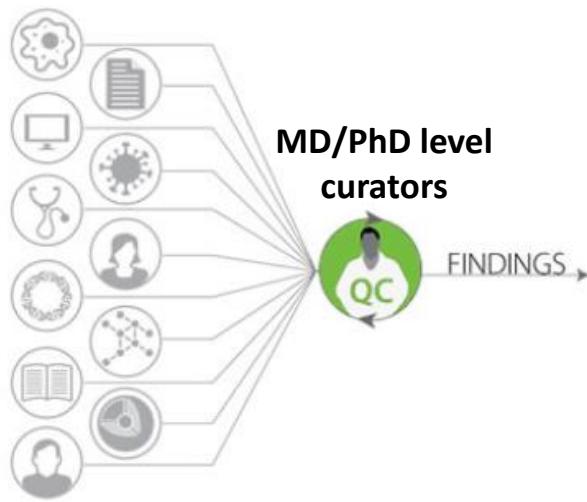


CLC Genomics Workbench + Ingenuity Pathway Analysis

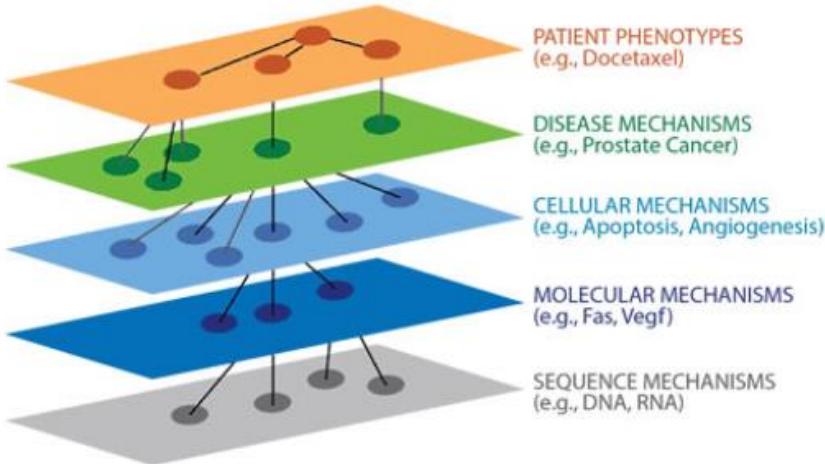
15年もの間メンテナンスされているナレッジベース

Ingenuity ナレッジベースはタンパク、遺伝子、複合体、細胞、組織、薬、疾患に関する数百万にも及ぶ生物学的機能、相互作用に関する情報が収集されたデータベースです。

Literature findings



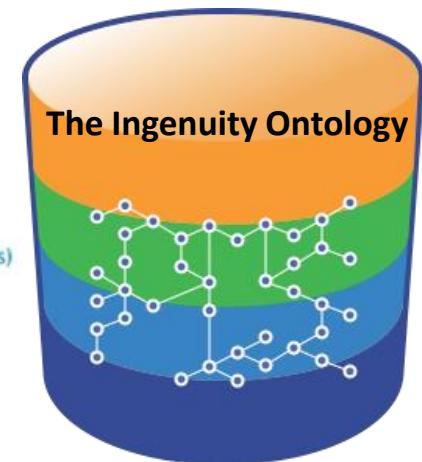
Biomedical Ontology



Content Acquisition

Ingenuity Ontology

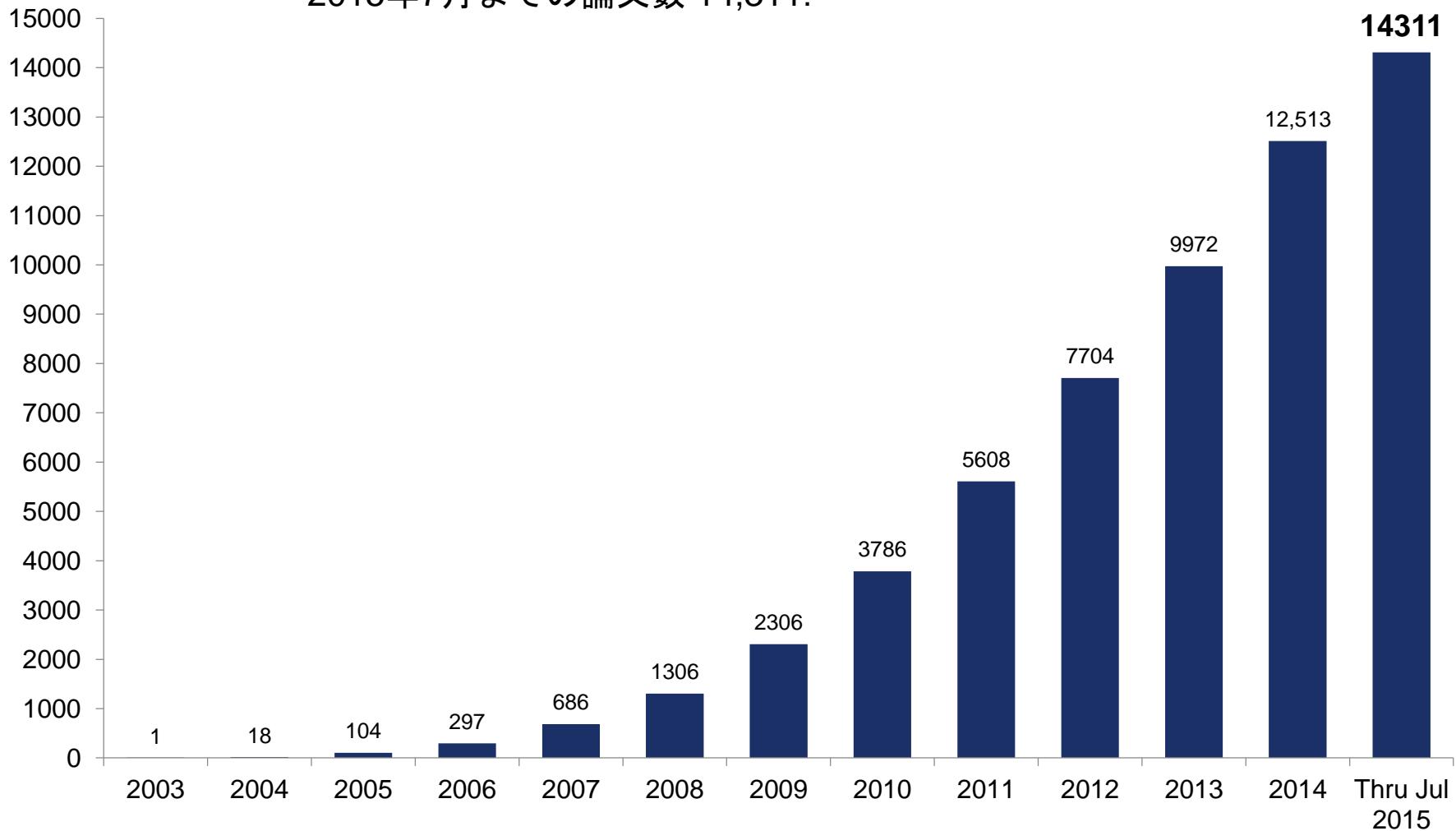
The Ingenuity Knowledge Base



POWERED BY
INGENUITY®

Ingenuity Pathway Analysis を使った論文数

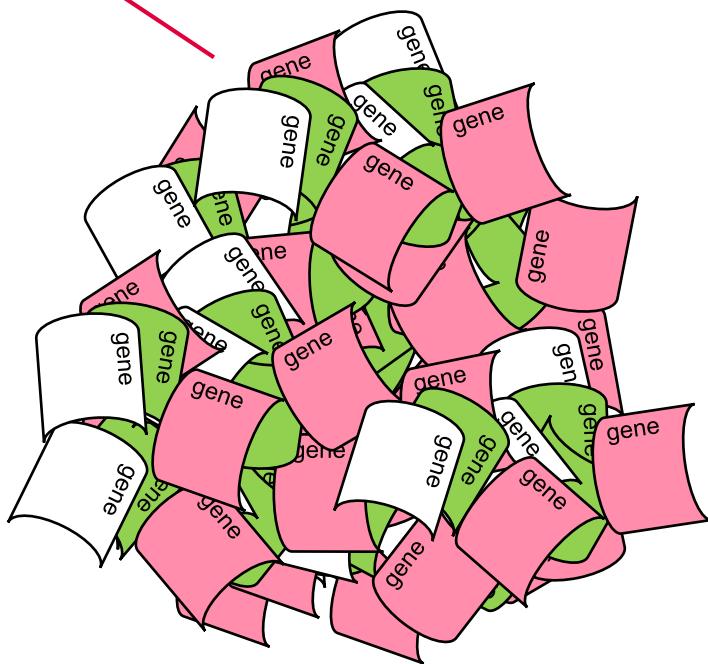
2015年7月までの論文数 14,311!



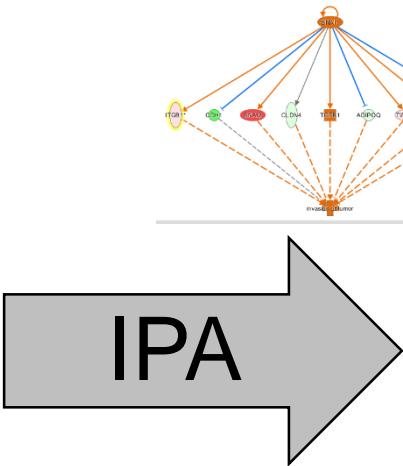


IPA?

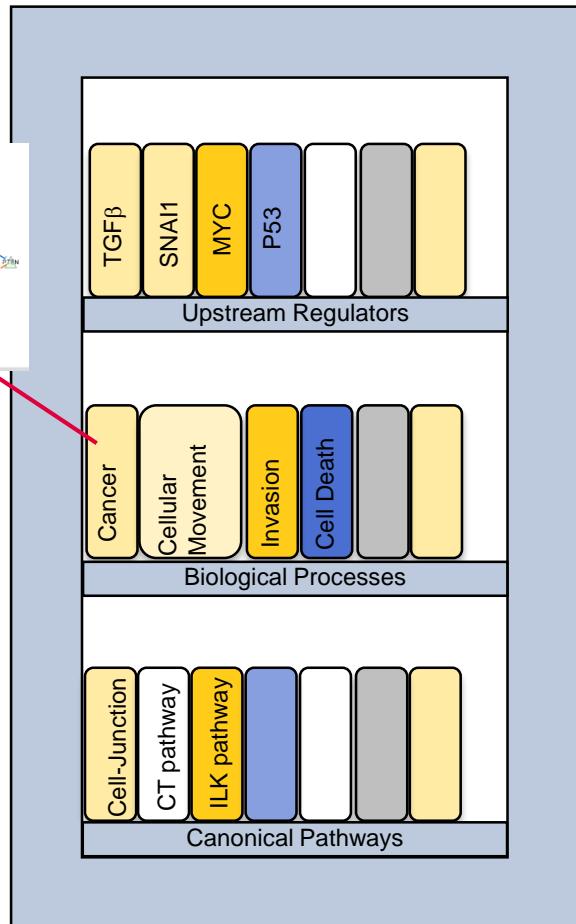
	A	B	C	D
ID		LogRatio	p-value	Intensity/PKRFM/KKM
1	NR_130786	0.14	6.86E-01	2319 69
3	NR_151380	-2.02	2.24E-01	1649 26
5	NR_151382	-0.02	0.95E-01	1649 26
5	NR_14576	0.02	0.85E-01	1.77
6	NR_138932	0.02	9.79E-01	1.83
7	NR_000014	-4.79	1.02E-01	239 75
8	NR_026971	-0.67	6.17E-01	213 79
9	NR_144670	-5.96	1.30E-01	610 64
10	NR_001080438	-1.97	3.47E-01	3.91
11	NR_017436	-1.05	5.02E-01	6168 93
12	NR_016161	-0.20	5.02E-01	149 85
13	NR_015565	0.27	3.68E-01	13334 90



RNA-seqからリストアップした遺伝子。論文を見ても山のような情報・・・



IPA



IPAヘリストをUp、Downの情報を送ることで、関連するパスウェイ、上流でコントロールしている遺伝子、表現型との関連や予測も行える

Core Analysis Summary

サマリーページでどのようなパスウェイの関連が強く考えられるか、Regulatorはどのような可能性があるか、一覧できる。

The screenshot shows the QIAGEN Sample to Insight interface with the following sections:

- Top Canonical Pathways:**

Name	p-value	Ratio
tRNA Charging	3.64E-07	9/39 (0.231)
Calcium Signaling	9.87E-06	16/178 (0.09)
Neuropathic Pain Signaling In Dorsal Horn Neurons	3.82E-05	11/100 (0.11)
Glutamate Receptor Signaling	7.75E-05	8/57 (0.14)
Breast Cancer Regulation by Stathmin1	8.9E-05	15/191 (0.079)
- Top Upstream Regulators:**

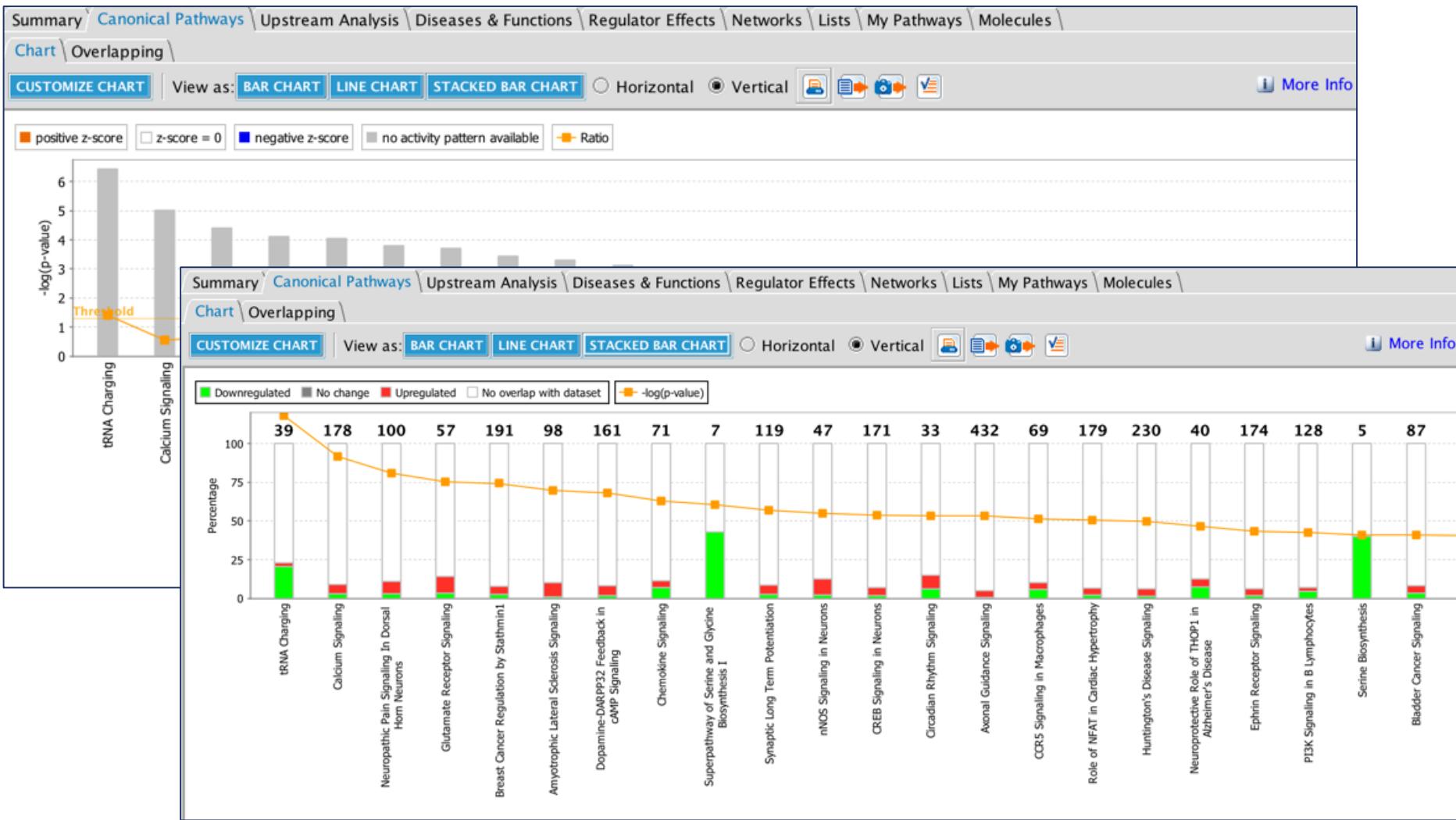
Upstream Regulator	p-value of overlap	Predicted Activ...
ATF4	2.51E-23	Inhibited
HTT	7.53E-21	Inhibited
APP	1.31E-18	
EIF2AK3	1.12E-16	Inhibited
NGF	5.54E-16	
- Causal Networks:**

	p-value of overlap	Predicted Activ...
S100A1	6.79E-23	
ketanserin	1.98E-21	
RGN	6.27E-21	
ladostigil	8.12E-21	
ADCYAP1R1	2.90E-20	
- Top Diseases and Bio Functions:**

Diseases and Disorders	p-value	# Molecules
Neurological Disease	2.59E-18 - 9.28E-04	192
Psychological Disorders	9.19E-14 - 9.22E-04	115
Hereditary Disorder	1.46E-12 - 8.71E-04	112

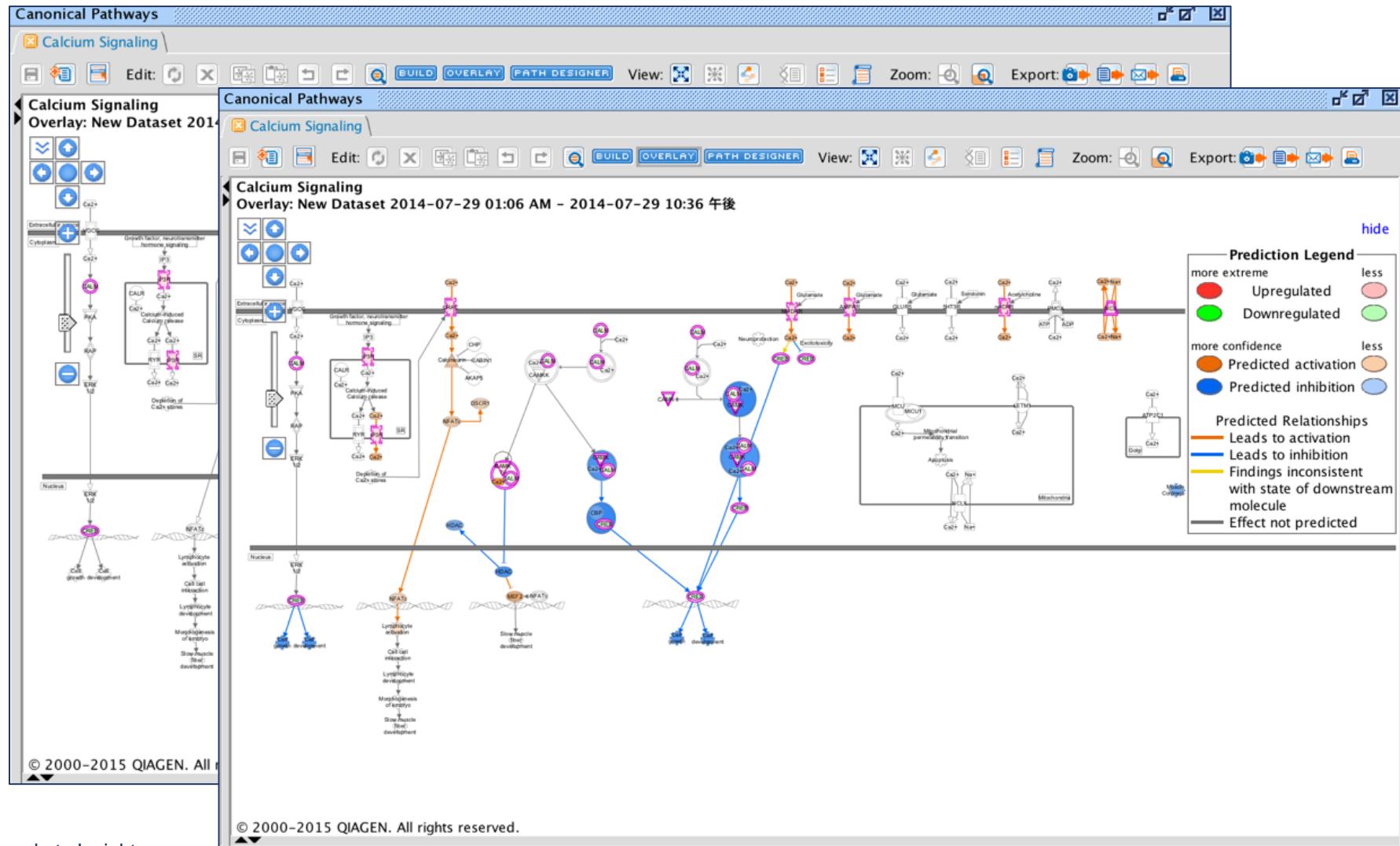
Canonical Pathway Analysis

関連すると考えられるパスウェイとp-valueをスコア化した値で閲覧



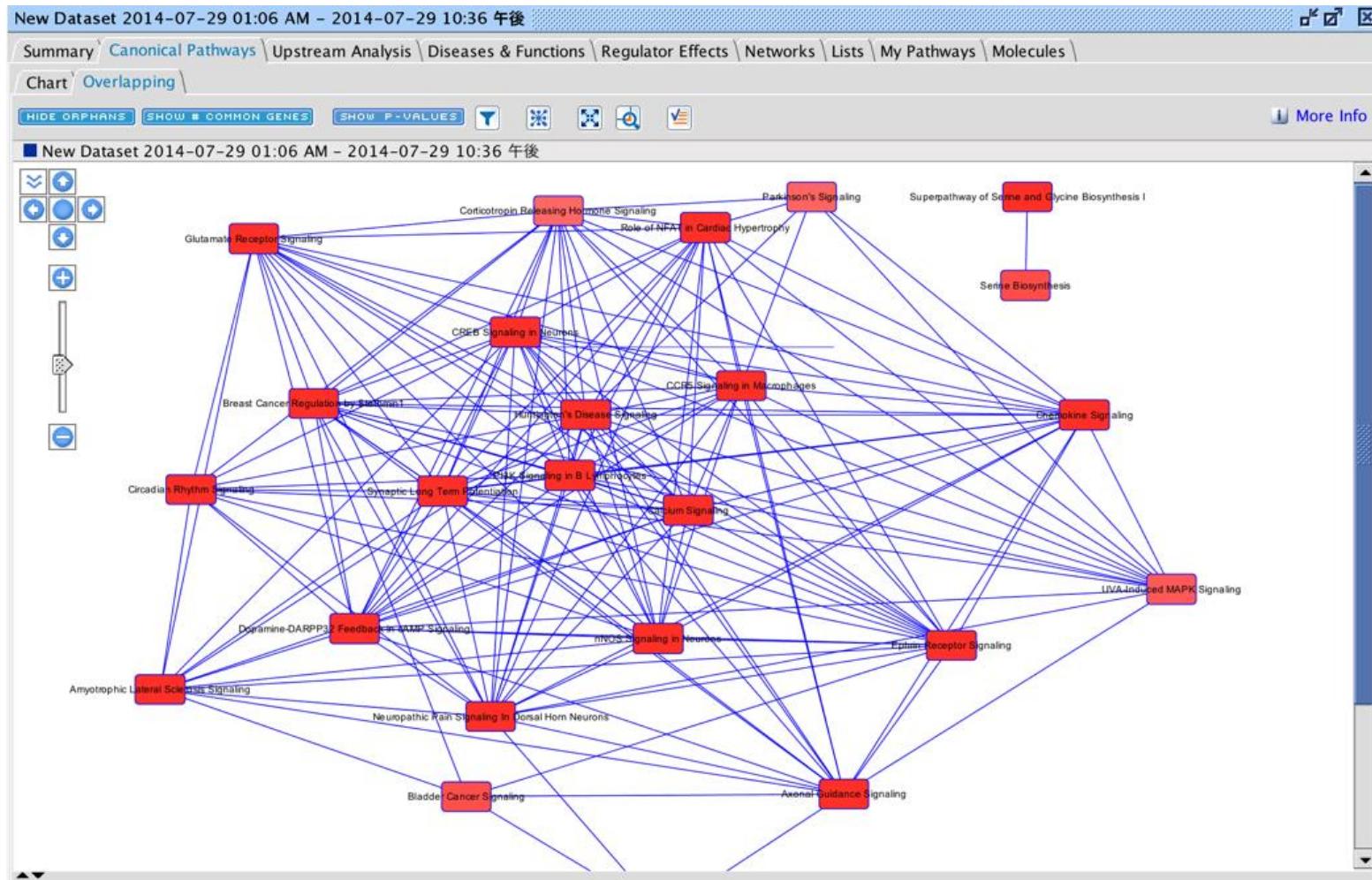
Canonical Pathway Analysis

リストに含まれている遺伝子はパスウェイ上でハイライト、Overlayツールで制御関係も重ね書き



Canonical Pathway

パスウェイ同士のつながりを表示。



見慣れない遺伝子も

その遺伝子に関する詳細な情報をリンクをクリックするだけで確認出来る。

IPA Gene View: ATF4 (Mammalian) > Interaction Network > View Reagents (102)

Review the categorized literature findings and database information for this node.

Summary Human Mouse Rat

Member Of: Atf, Creb

Entrez Gene Name: activating transcription factor 4

Synonym(s): activating transcription factor 4, C/ATF, CREB-2, TAXREB67, TXREB

NCBI CDD Domains (Superfamilies / Multi-Domains): bZIP

Protein Functions / Functional Domains: activation domain, basic amino acid motif, basic domain, DNA binding, DNA binding domain, leucine zipper domain, protein binding, protein C-terminus binding, transcription activation domain, transcription factor, transcription regulator

Subcellular Location: cell borders, chromatin, Cytoplasm, dendrites, intracellular space, neurites, nuclear fraction, nucleoplasm, Nucleus, perikaryon, perinuclear region

Canonical Pathway: ATM Signaling; B Cell Receptor Signaling; Calcium Signaling; cAMP-mediated signaling; Circadian Rhythm Signaling; Corticotropin Releasing Hormone Signaling; CREB Signaling in Neurons; Dendritic Cell Maturation; Dopamine-DARPP32 Feedback in GMP Signaling; Endoplasmic Reticulum Stress Pathway; Ephrin Receptor Signaling; ERK/MAPK Signaling; ERK5 Signaling; Estrogen-Dependent Breast Cancer Signaling; FGF Signaling; FLT3 Signaling in Hematopoietic Progenitor Cells; GNRH Signaling; G-Protein Coupled Receptor Signaling; Gα Signaling; Huntington's Disease Signaling; Hypoxia Signaling in the Cardiovascular System; ILK Signaling; Melanocyte Development and Pigmentation Signaling; Neurotrophin/TRK Signaling; NGF Signaling; NRF2-mediated Oxidative Stress Response; P2Y Purinergic Receptor Signaling Pathway; p38 MAPK Signaling; Phospholipase C Signaling; PI3K Signaling in B Lymphocytes; Prostate Cancer Signaling; Protein Kinase A Signaling; Role of IL-17F in Allergic Inflammatory Airway Diseases; Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis; Synaptic Long Term Potentiation; Unfolded protein response; Wnt/Ca²⁺ pathway

Targeted By miRNA Functional Cluster: miR-214-3p (and other miRNAs w/seed CAGCAGG), miR-3162-3p (miRNAs w/seed CCCUACC), miR-4446-5p (miRNAs w/seed UUUCCC), miR-548v (miRNAs w/seed GCUCACG)

Top findings from Ingenuity Knowledge Base (show all 1444 categorized literature findings)

- regulates:** DDIT3, ATF3, ASNS, HERPUD1, VEGFA, SLC38A2, PPP1R15A, CEBPB, SARS, WARS, SLC3A2, MAP1LC3B, HSPA5, HSP90B1, PCK2
- regulated by:** thapsigargin, L-histidine, D-glucose, dexamethasone, N-acetyl-L-cysteine, homocysteine, bortezomib, benzyloxycarbonyl-Leu-Leu-Leu aldehyde, FGF2, tretinoin, NFE2L2, palmitic acid, EIF2AK3, LY294002, sirolimus
- binds:** JUN, UBC, FOS, DDIT3, DAPK3, BTRC, GABBR1, NFE2L2, TRIB3, CEBPA, ATF3, CREBBP, DISC1, GABBR2, EP300
- role in cell:** expression in, apoptosis, endoplasmic reticulum stress response in, proliferation, cell death, gluconeogenesis in, survival, transcription in, abnormal morphology, activation in
- disease:** microphthalmia, hypoplasia, growth failure, eyelids fail to open, aphakia, anemia, hepatic steatosis, hypertriglyceridemia, Parkinson's disease, endometriosis

Human Isoforms From RefSeq : More Info

ATF4 Chromosome: 22; Location: 22q13.1

Descriptions from External Databases

Entrez Gene Summary: This gene encodes a transcription factor that was originally identified as a widely expressed mammalian DNA binding protein that could bind a tax-responsive enhancer element in the LTR of HTLV-1. The encoded protein was also isolated and characterized as the cAMP-response element binding protein 2 (CREB-2). The protein encoded by this gene belongs to a family of DNA-binding proteins that includes the AP-1 family of transcription factors, cAMP-response element binding proteins (CREBs) and CREB-like proteins. These transcription factors share a leucine zipper region that is involved in protein-protein interactions, located C-terminal to a stretch of basic amino acids that functions as a DNA binding domain. Two alternative transcripts encoding the same protein have been described. Two pseudogenes are located on the X chromosome at q28 in a region containing a large inverted duplication. [provided by RefSeq, Sep 2011]

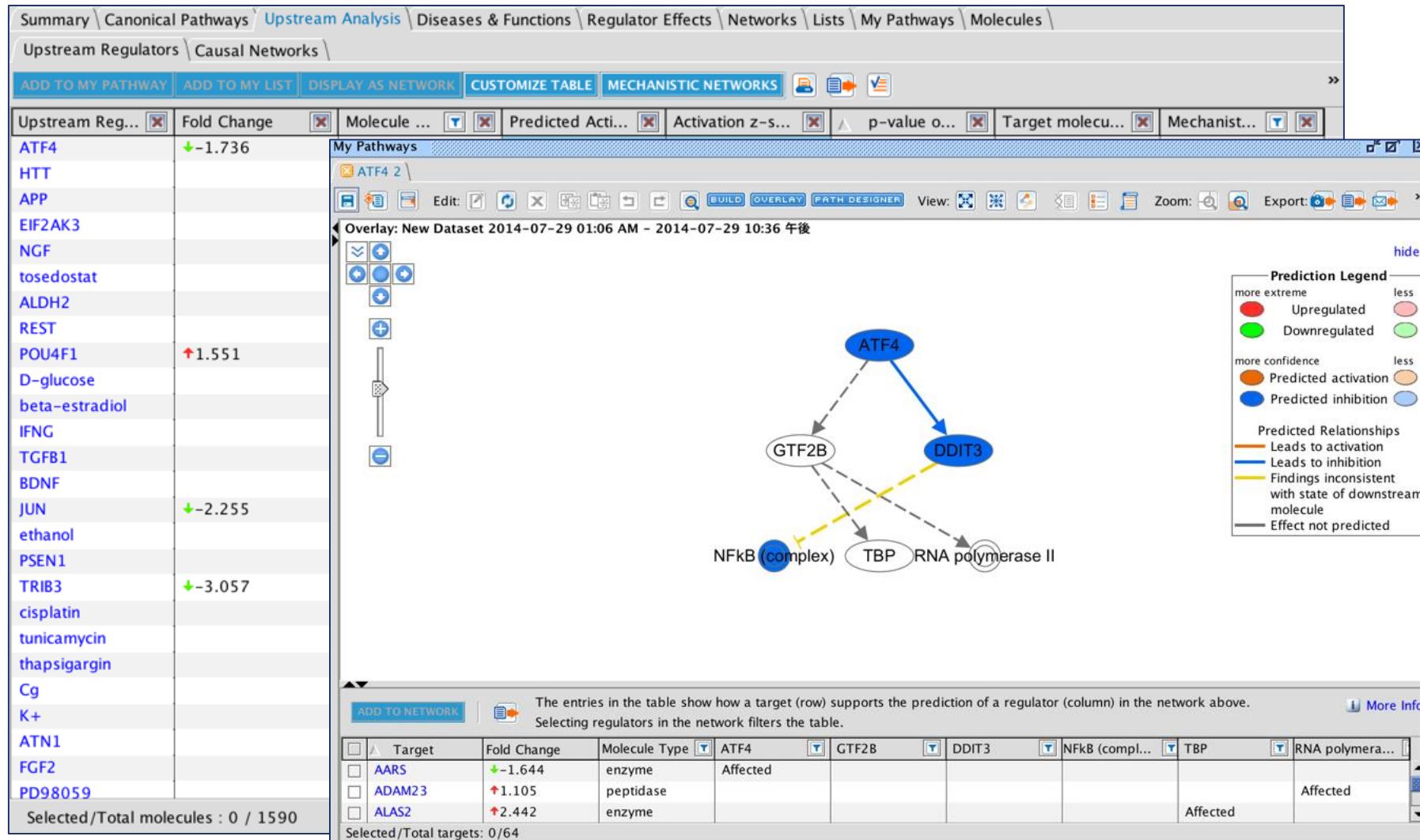
GO Annotations

Molecular Function: core promoter sequence-specific DNA binding; DNA binding; protein binding; protein C-terminus binding; RNA polymerase II core promoter proximal region sequence-specific DNA binding; RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription; sequence-specific DNA binding; sequence-specific DNA binding RNA polymerase II transcription factor activity; sequence-specific DNA binding transcription factor activity; transcription regulatory region DNA binding

Biological Process: activation of signaling protein activity involved in unfolded protein response; cellular protein metabolic process; circadian regulation of gene expression; endoplasmic reticulum unfolded protein response; gamma-aminobutyric acid signaling pathway; gluconeogenesis; intrinsic apoptotic signaling pathway in response to endoplasmic reticulum stress; negative regulation of potassium ion transport; positive regulation of apoptotic process; positive regulation of neuron apoptotic process; positive regulation of transcription, DNA-dependent; positive regulation of transcription from RNA polymerase II promoter; positive regulation of transcription from RNA polymerase II promoter in response to oxidative stress; regulation of transcription, DNA-dependent; regulation of transcription from RNA polymerase II promoter; response to endoplasmic reticulum stress; transcription, DNA-dependent; transcription from RNA polymerase II promoter

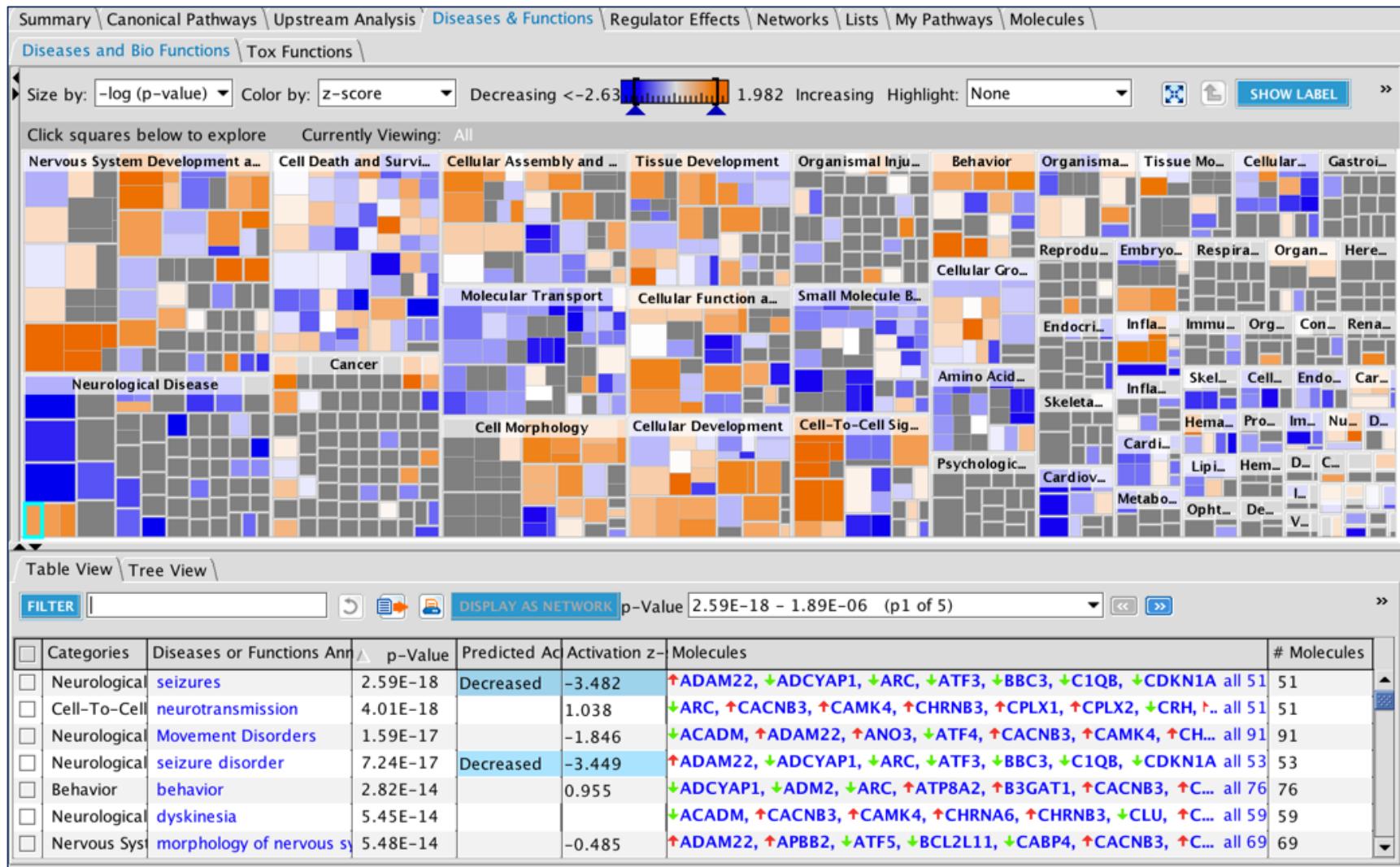
Upstream Analysis

データより上流で調節をしていると思われるRegulatorを予測した結果



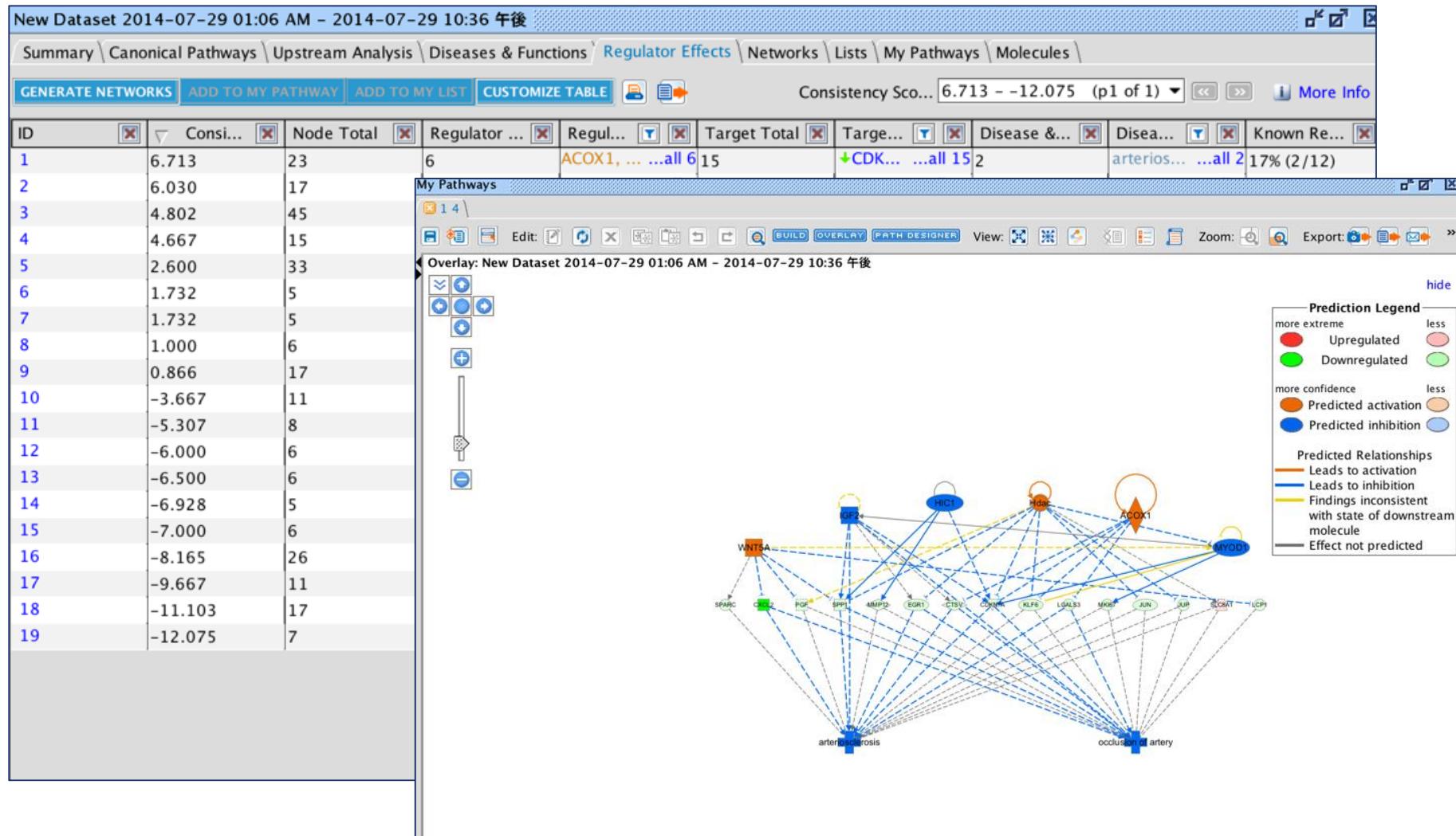
Disease and Function

アップロードした遺伝子のリストは、どういった疾患や機能に関連しているのか

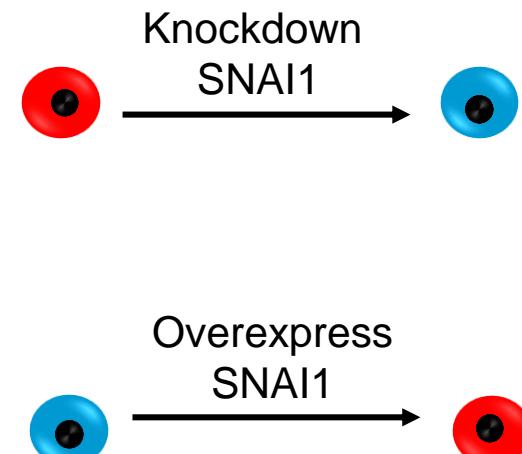
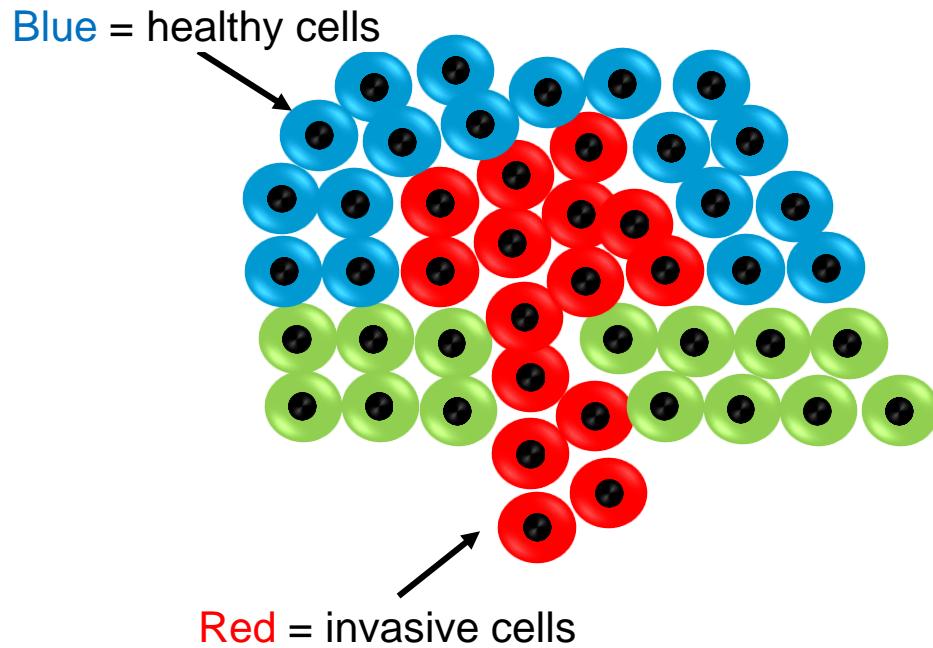


Regular Effect

Upstream Regulator が機能や疾患にどのような影響を与えるのか



Case Study: SNAI1ノックダウンと癌の浸潤性

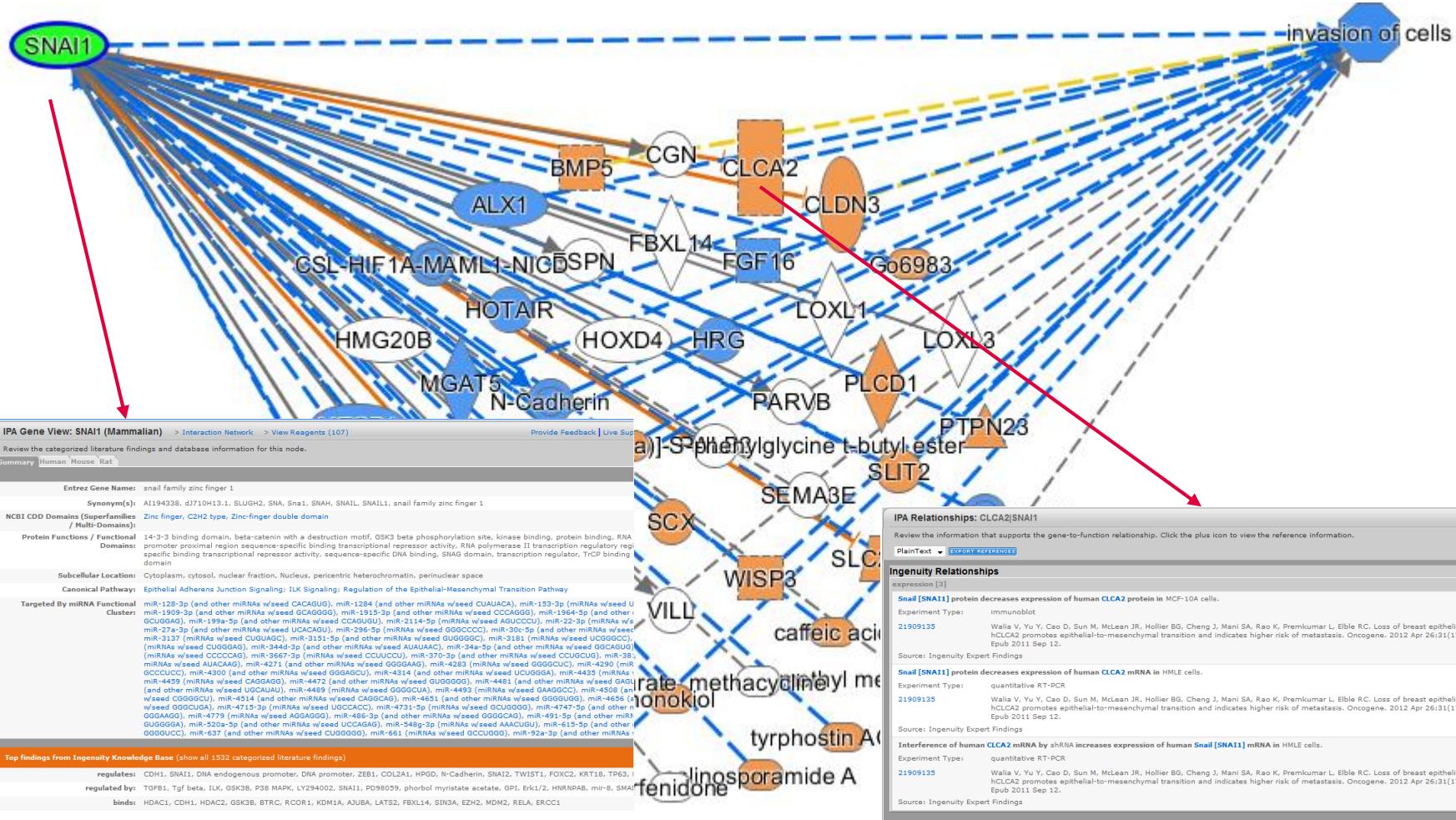


Case Study: SNAI1ノックダウンと癌の浸潤性

My Pathways

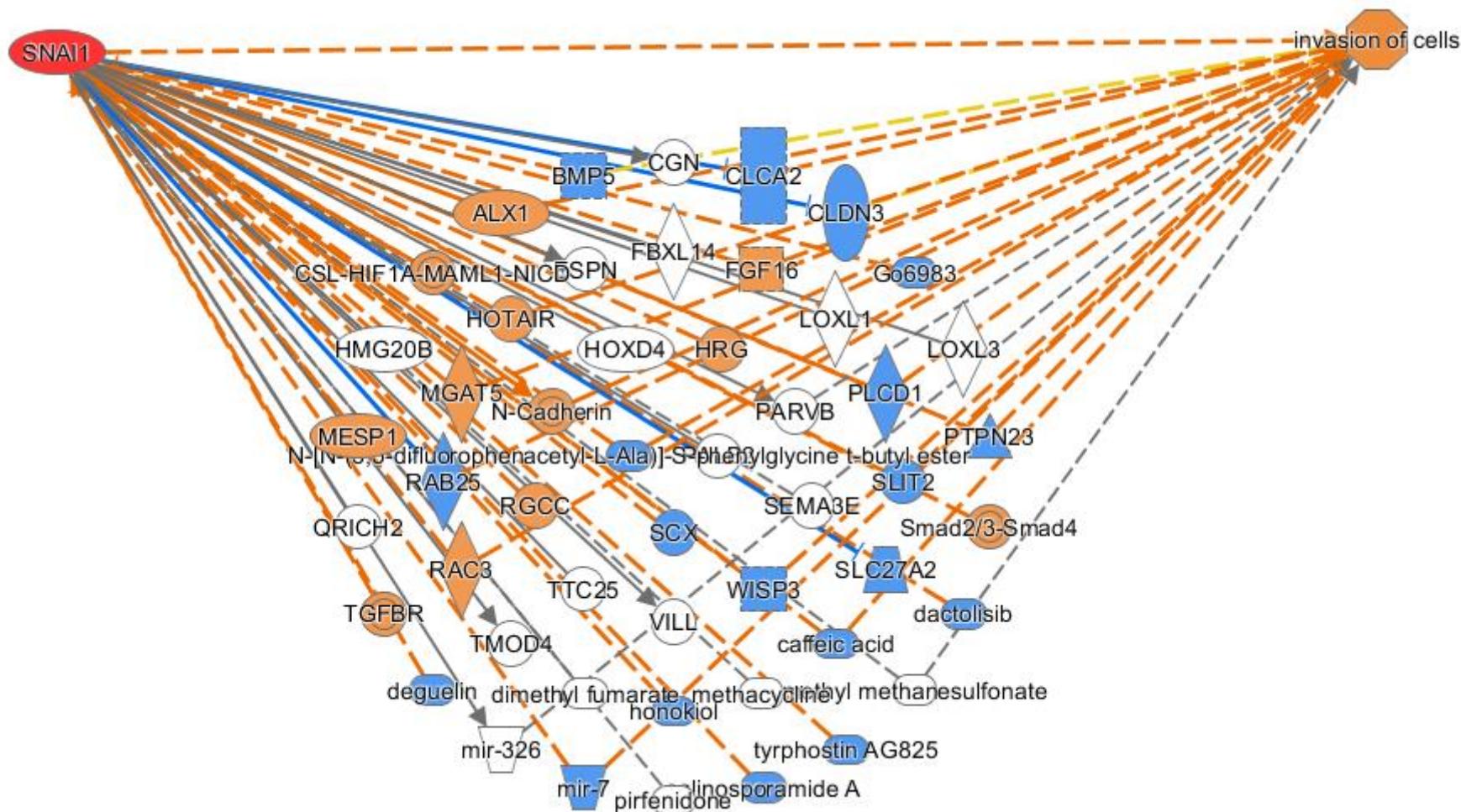
New My Pathway 5

Edit: BUILD OVERLAY PATH DESIGNER View: Zoom: Export:



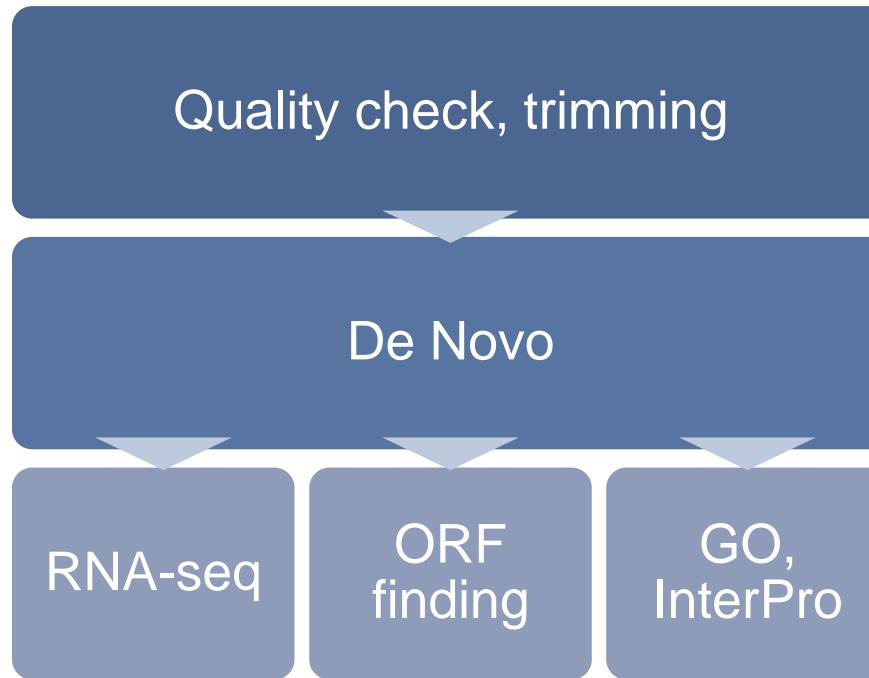
Sample to Insight

Case Study: SNAI1 knockdown making cells less invasive



CLC Genomics Workbench + BLAST2GO plug-in

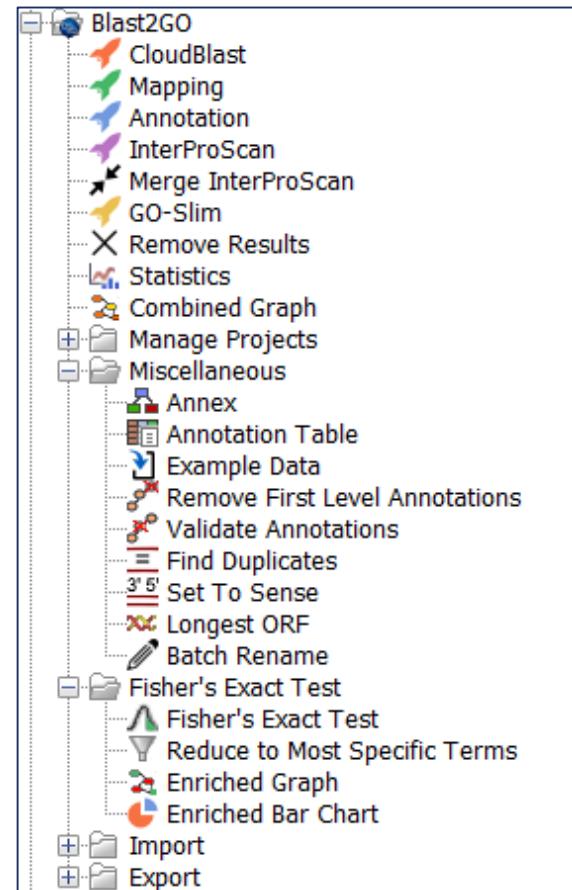
非モデル生物の場合



プラグインを使い、機能拡張

BLAST2GO プラグイン

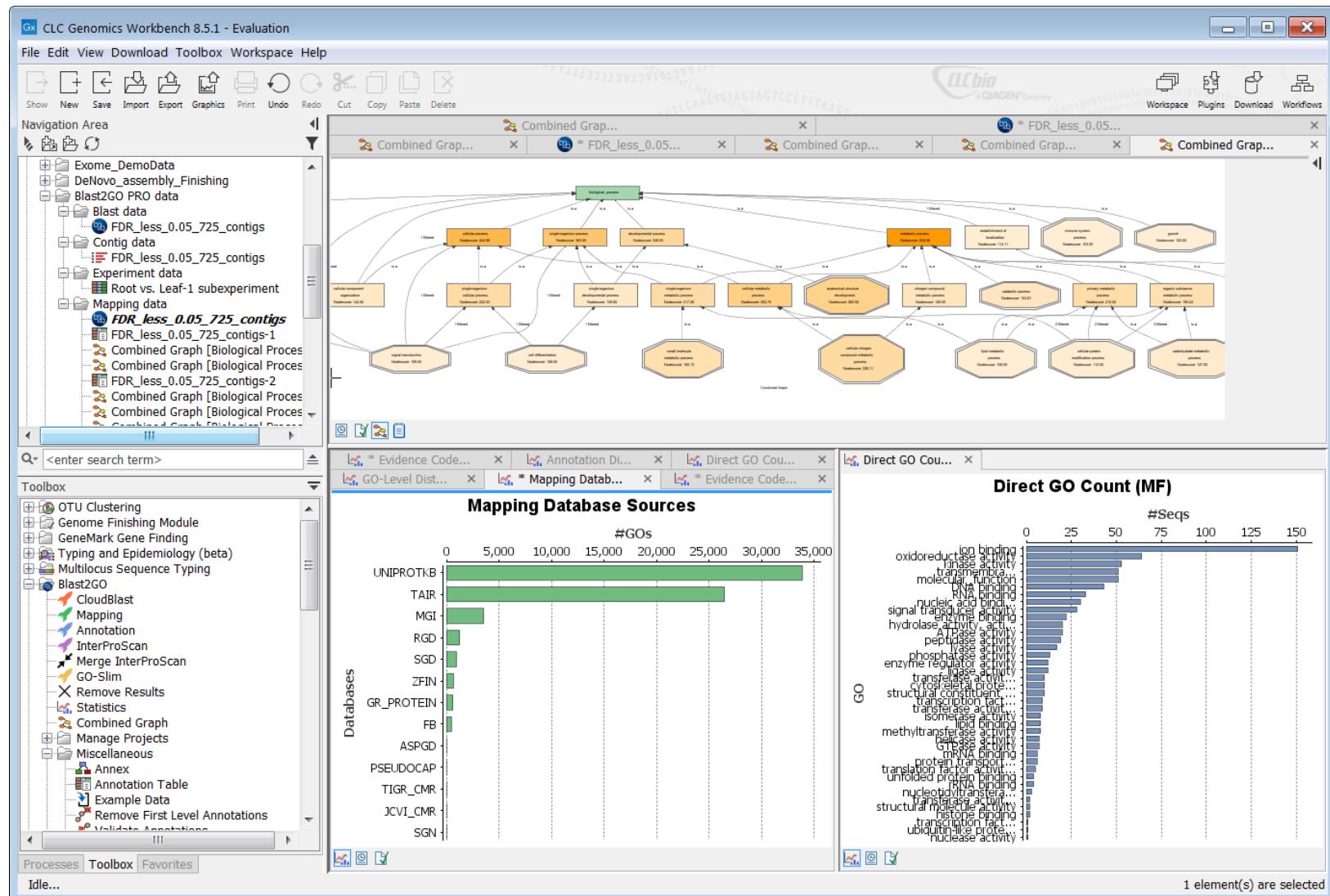
- BLAST2GO プラグインはBioBam社が作成しているBLAST2GO PROをWorkbenchで実行できるように作成されたプラグインです。
- 非モデル生物など、ゲノム配列の解読が行われておらず、アノテーションが未知の生物を研究する場合、研究者自身でのアノテーション付けが必要となります。
- BLAST2GOでは、配列をBLASTにかけ、GOのアノテーション付け、さらにInterProScanによる配列モチーフ、ドメインに基づいたアノテーションの取得が可能です。
- またアノテーション情報の統計値やビジュアライゼーションも可能です。
- さらにCloudBLAST機能を使えば、クラウド上でBLASTを実施でき、手元に大きなコンピュータ資源がない場合に大変便利です。





BLAST2GO プラグイン

結果例



Question?



mari.miyanoto@qiagen.com

補足資料

CLC Genomics Workbench

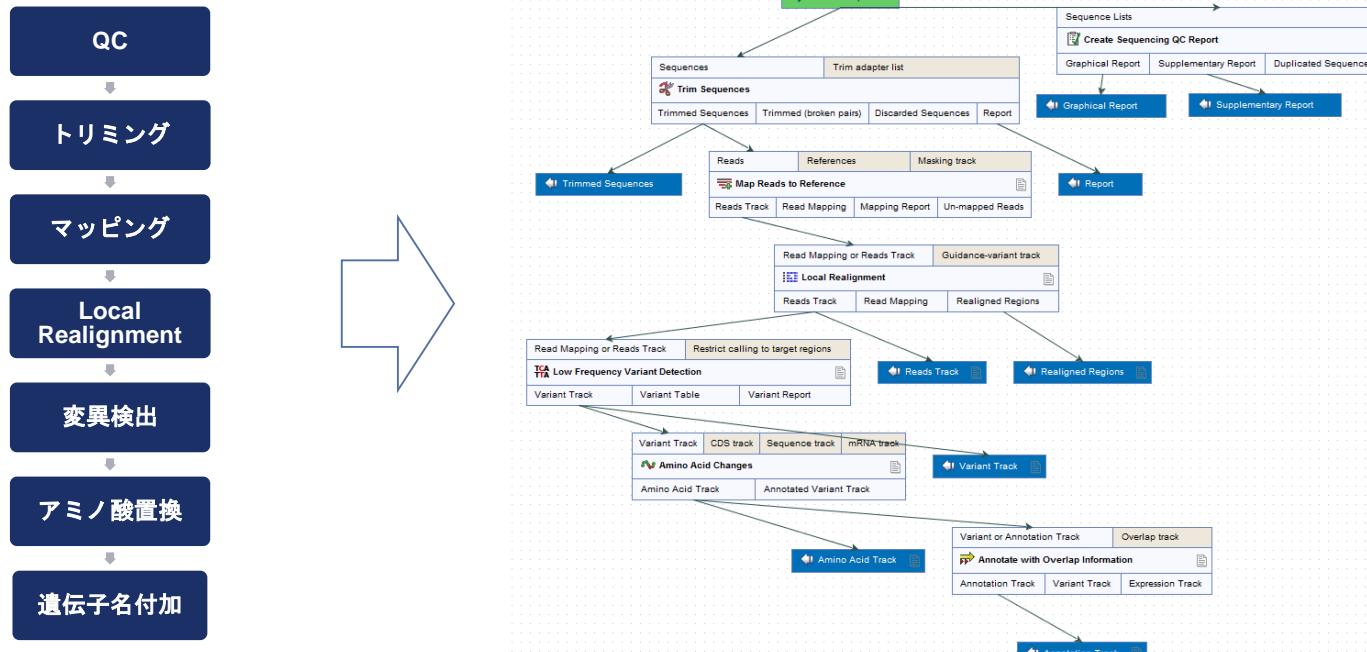
ワークフロー

ワークフローについて

一連の解析をひとつのフローに

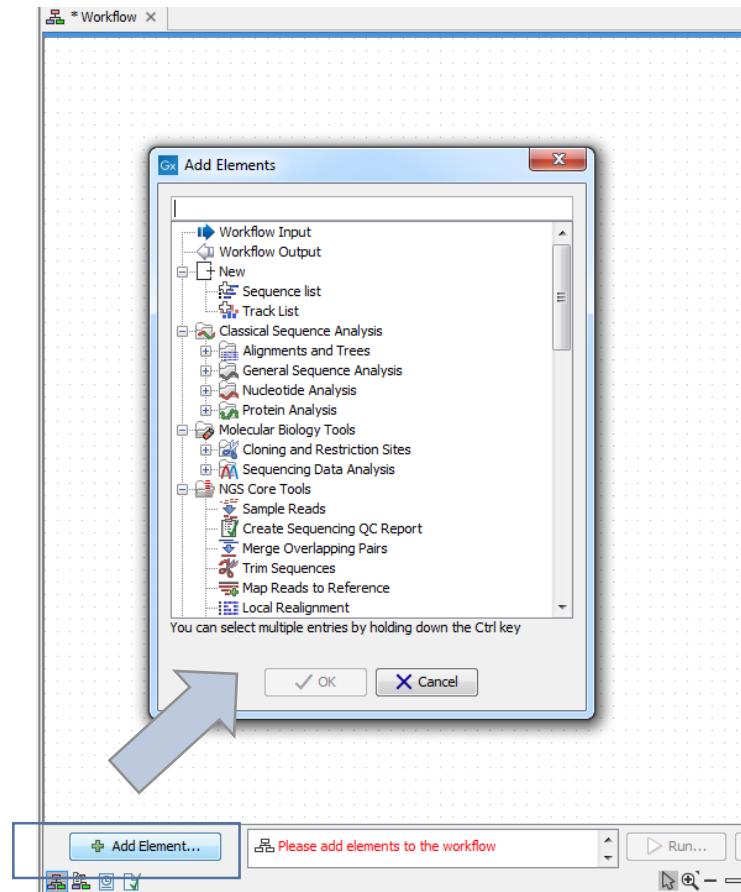
ワークフローツールは、様々な解析ツールを、フローチャートのようにつなげて、ひとつの解析のように実施することが可能です。

解析ツールをつなげてひとつの解析のようにすることを「解析パイプラインを作る」とも言いますが、Genomics Workbench の中では、解析パイプラインをワークフローと呼んでいます。



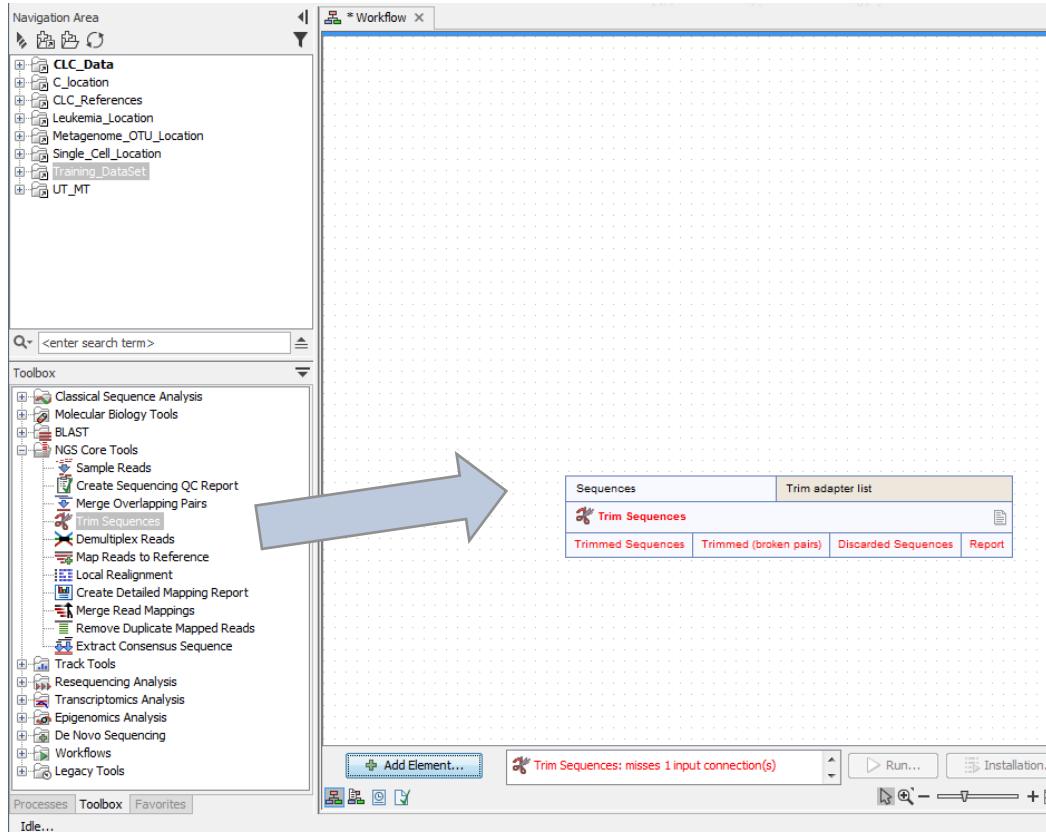
解析ツールの追加方法

2つの方法でツールを追加できます。
Add elements ボタンから追加



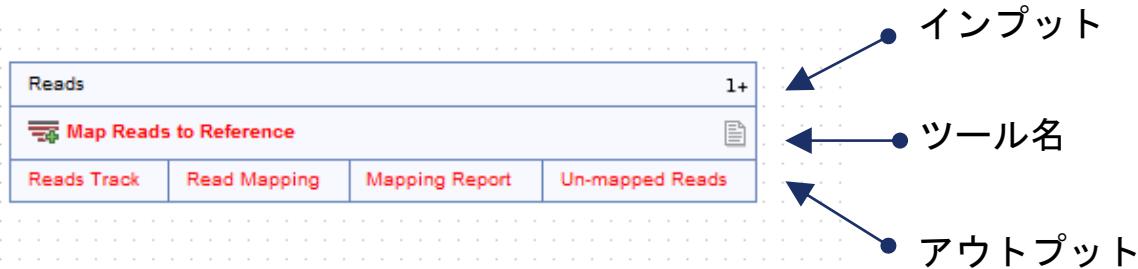
解析ツールの追加方法

Toolbox から ドラッグアンドドロップで追加

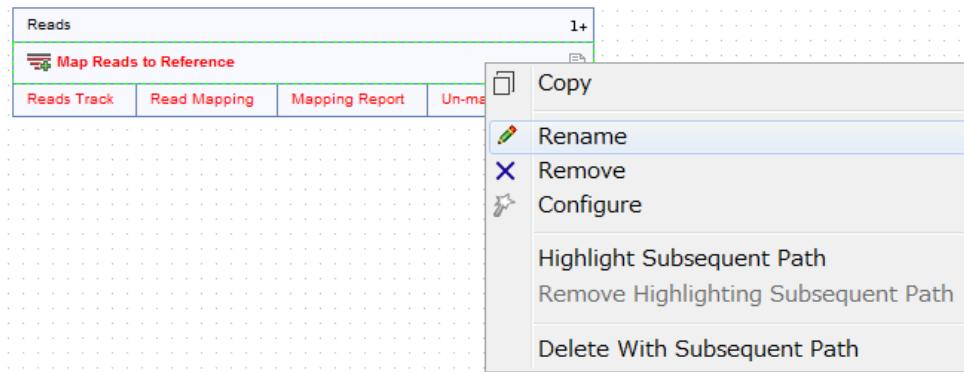


ツールの配置

※各ツールのことを workflow element と呼びます

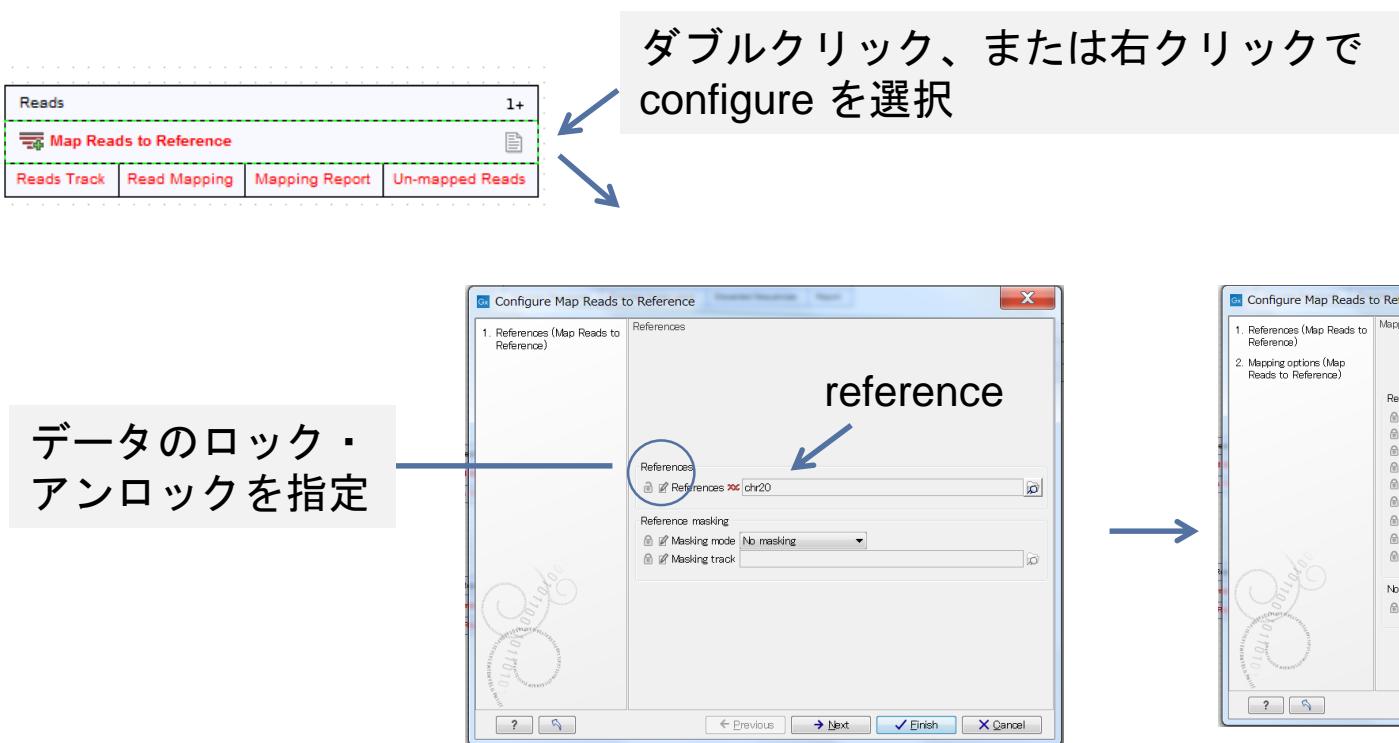


Name of the tool を rename して、
わかりやすい名前にすることができます



workflow element (ツール) のパラメータ設定

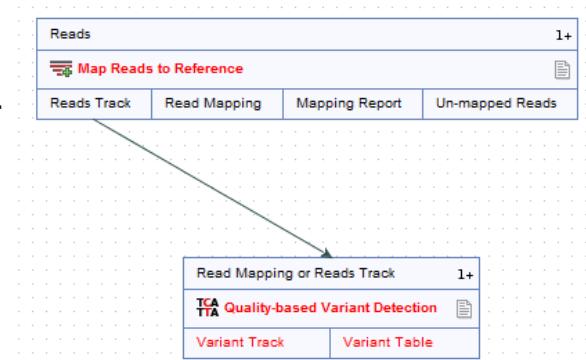
解析の際に必要な reference 配列やパラメータなどを設定します



Workflow : 作成

各 workflow element (ツール) の連結

次の解析で使う出力データと次の解析の Input box をつなぎます

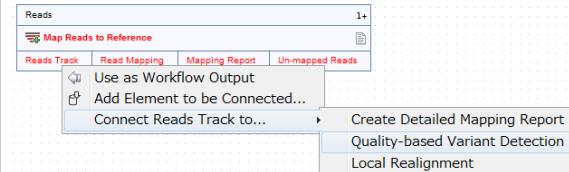


方法1.



出力データをドラッグして
次の解析の Input box
とつなぎます

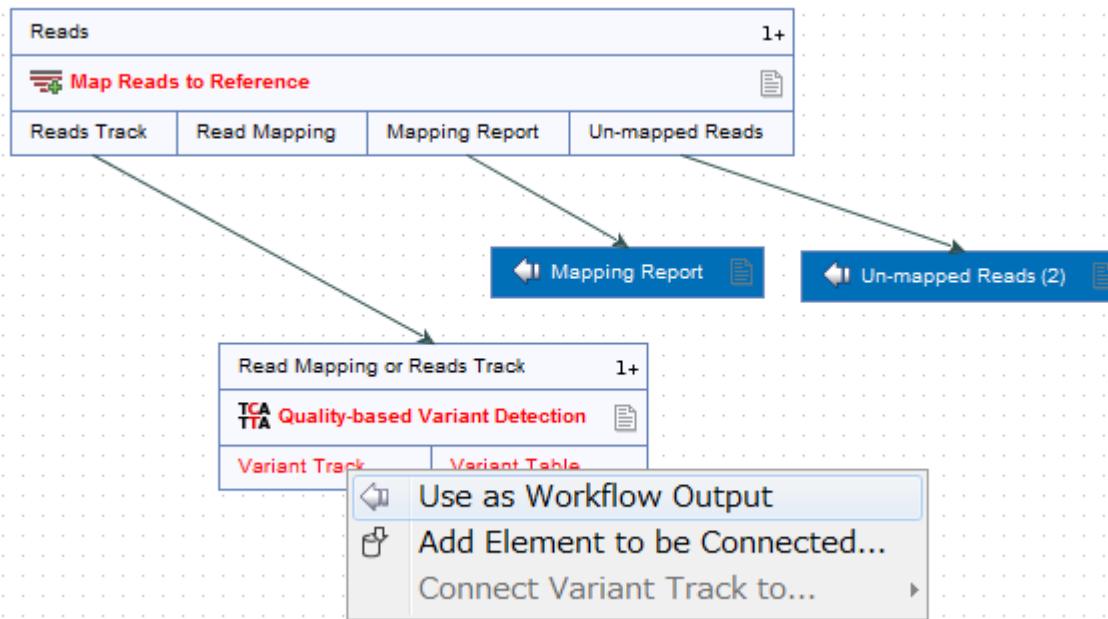
方法2.



出力データを右クリック
Connect Reads Track to...
から任意選択します

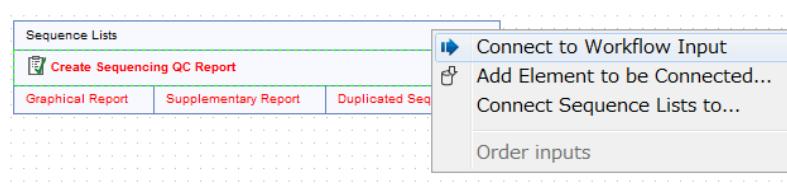
Output ファイルの指定

出力データを右クリックして Use as Workflow Outputで指定します。途中で出力・保存されるデータについても全て指定します（下図では、Mapping Report, Un-mapped Reads）



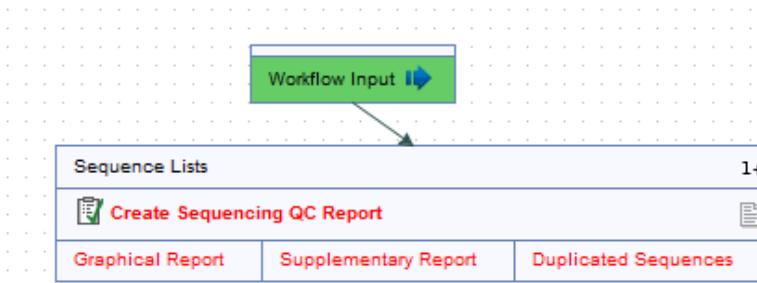
Input ファイルの指定

最初の入力データを受け取る workflow element (ツール) の Input box を右クリック、Connect to Workflow Input を選択します。



下図のような Workflow Input のボックスが作成されます。

このボックスをドラッグして、別のワークフローの最初の element とつなげることで、同じデータを複数のワークフローに対して渡して実行することができます。



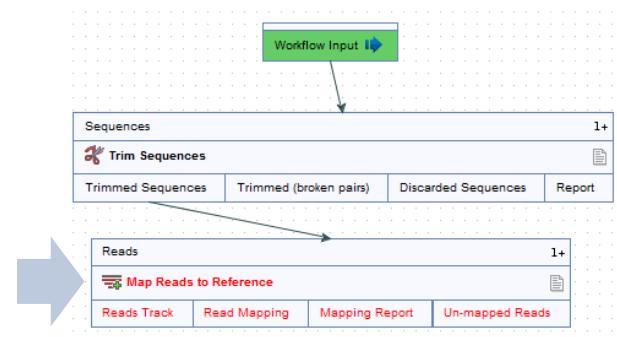
右クリックのメニューから Layout を選択すると自動的に Workflow が整列されます

Validation

- 必要な操作について表示されます



- 操作が必要な部分は赤字で表示されます



- 次のようなメッセージが表示されたら、Workflow を作成したら保存します

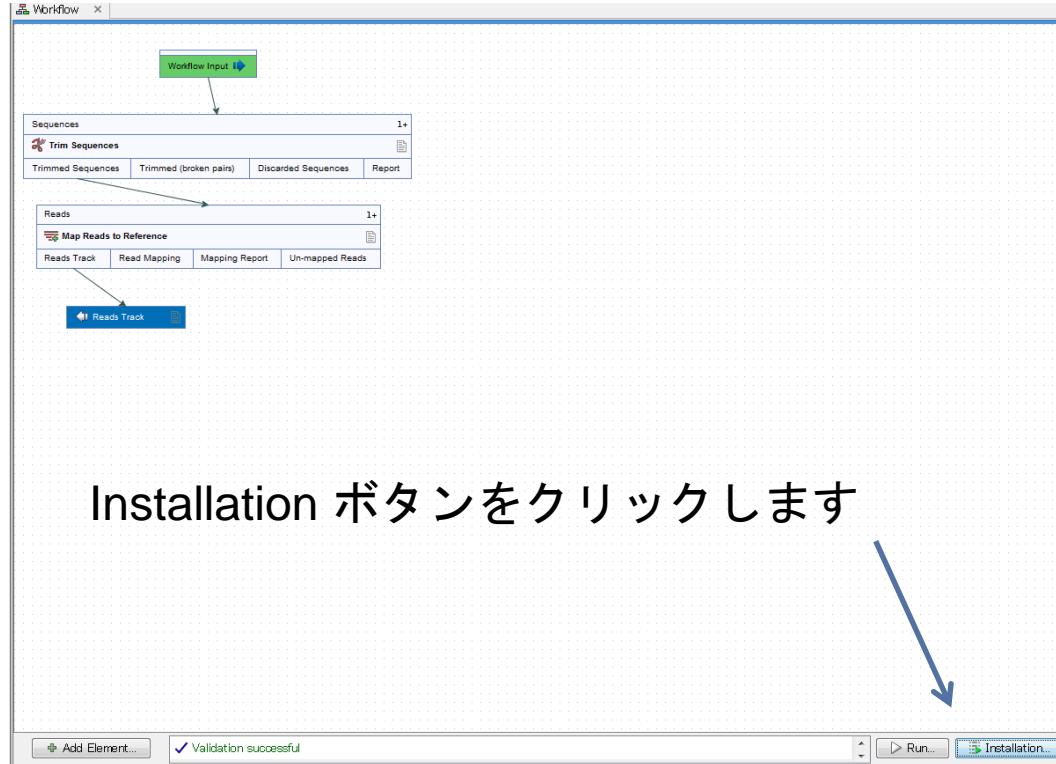


- Validation にパスした workflowでは、下記のようなメッセージが表示されます。



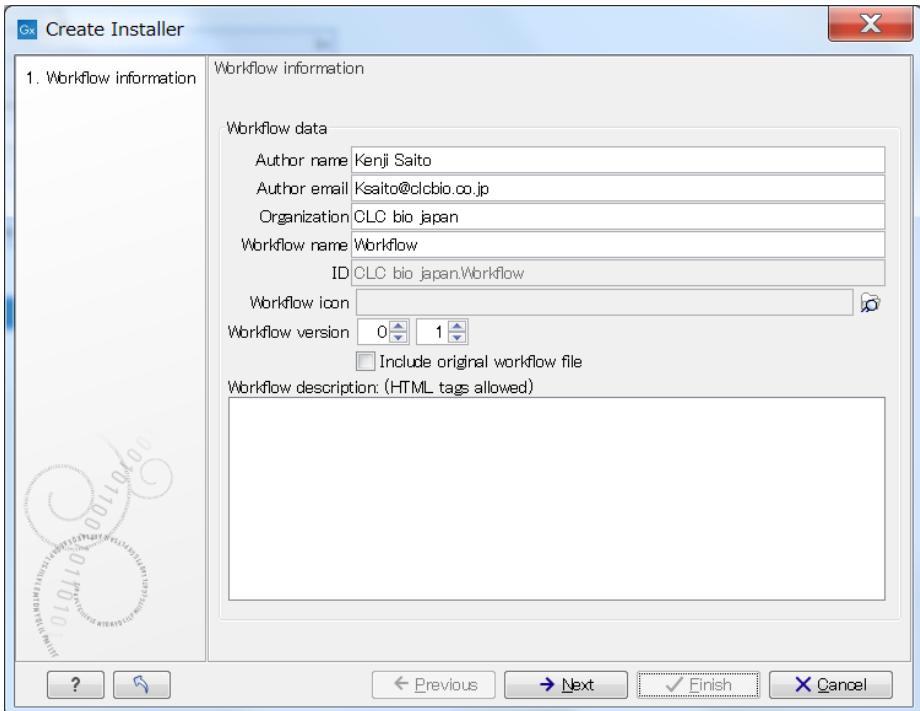
配布

Workflow Installation File を作成することにより、workflow を配布することができます。



配布

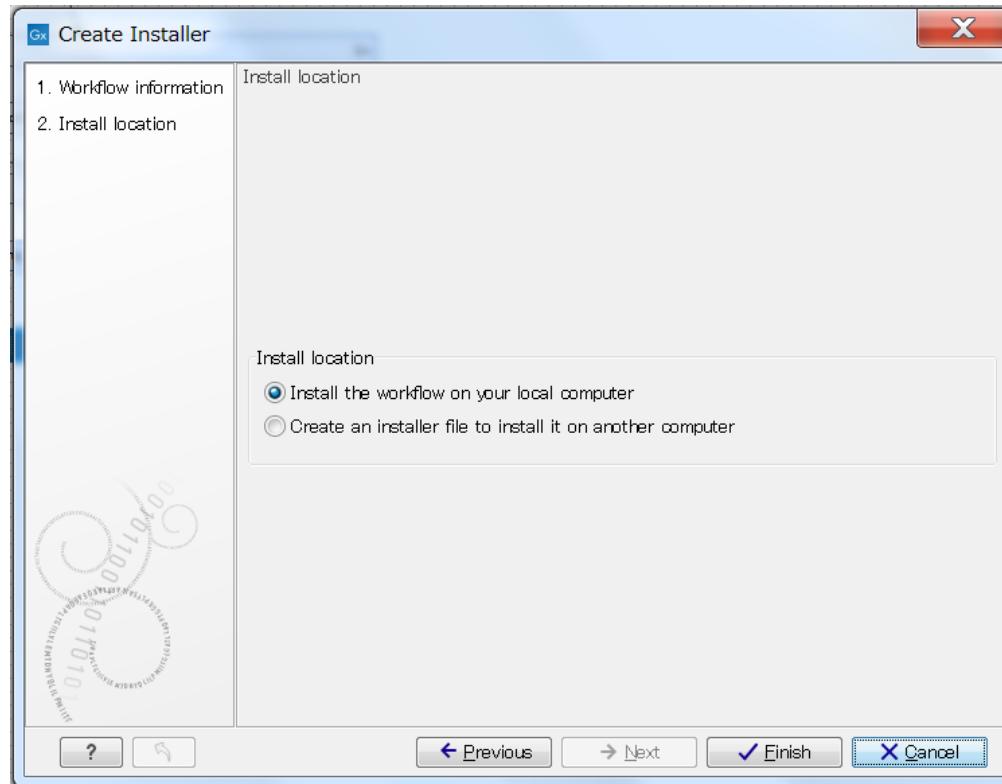
下記ダイアログに情報を入力します



- Author Information: 名前、メール、組織名
- Workflow Name: Workflow の名前
※workflow の IDは、組織名 + Workflowの名前から成ります。
- Workflow description: Workflow の説明
テキストまたは HTML (ver3.1互換)
- Icon (16 x 16 pixels gif or png)
- Version (major and minor)
- ※ 新しいバージョンを出したい時には、新たにバージョン情報を変えた Workflow Installation File を作成します。

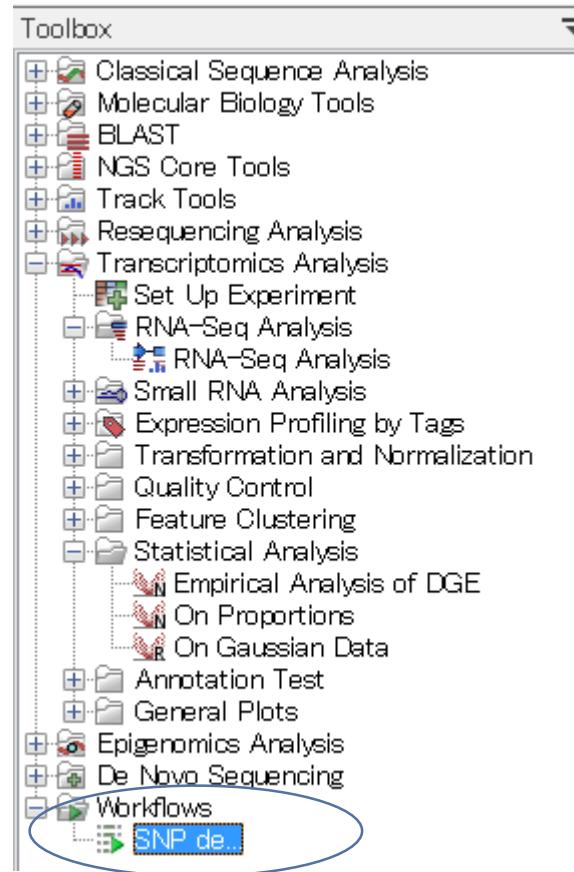
配布

workflow のインストール先のコンピュータまたは配布用ファイルの作成を選択します



実行

Toolbox の Workflow の下から workflow を選んで実行します



CLC Genomics Workbench

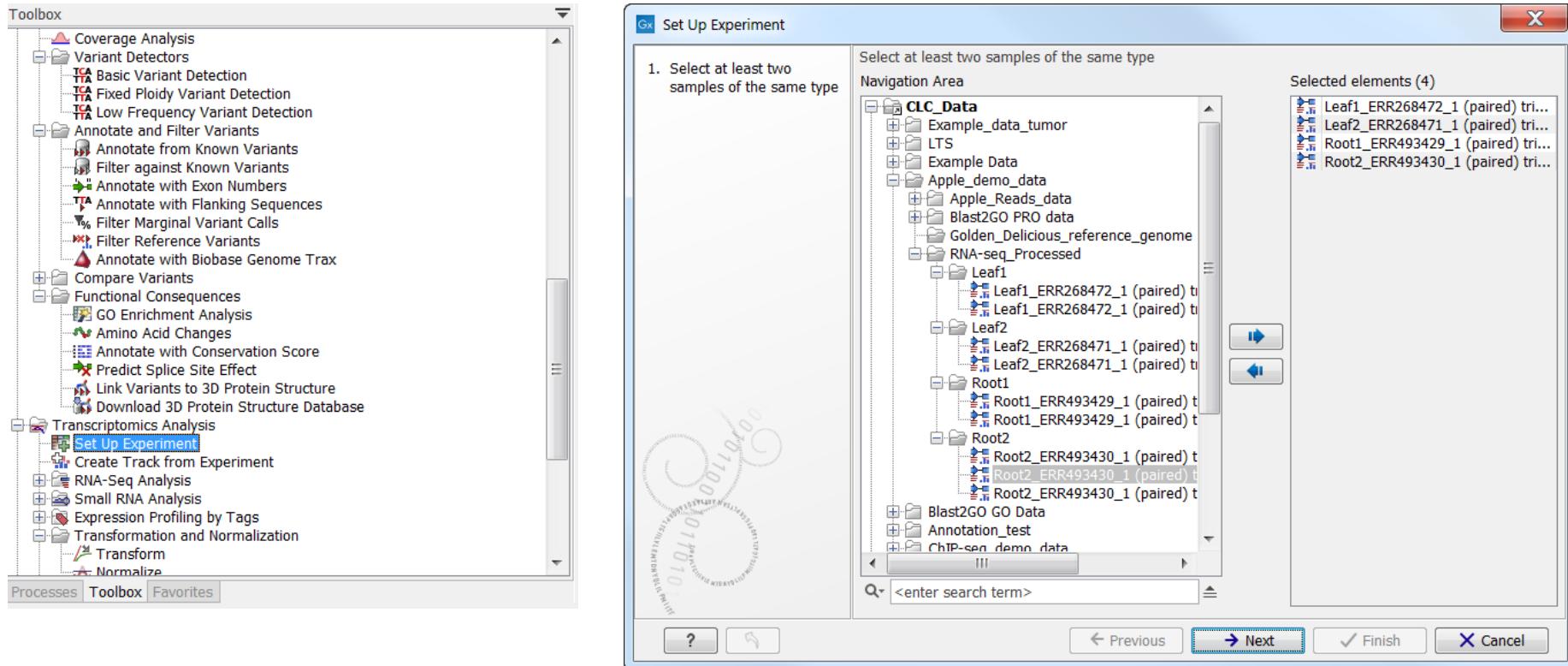
クラスタリング

全体の流れ



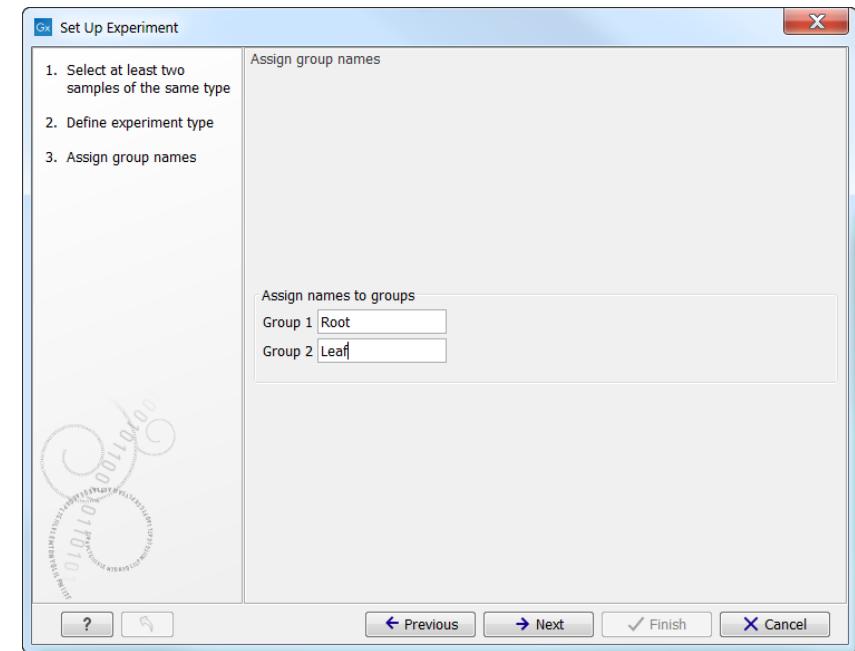
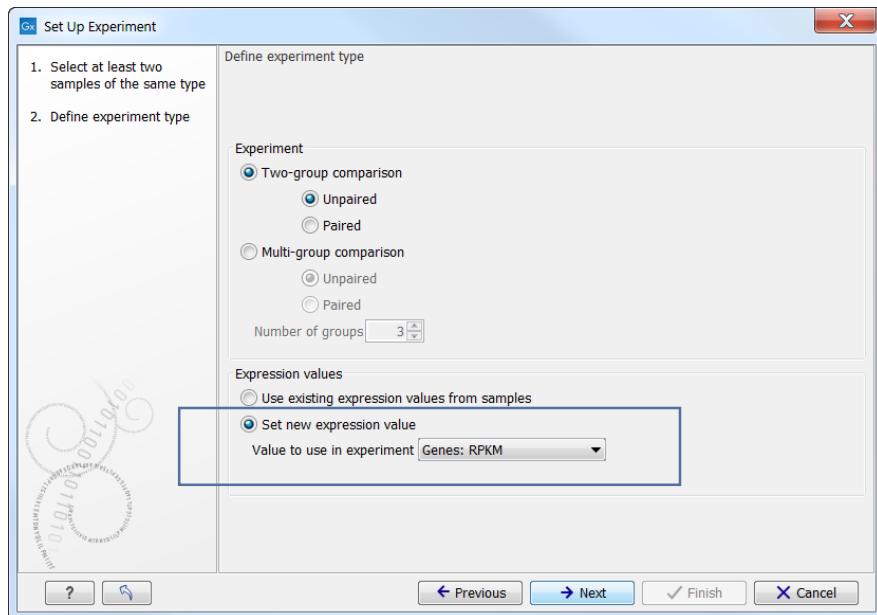
- クラスタリングでは発現差解析とは異なり、RPKMを使ってクラスタリングを行います。その際、ある程度の遺伝子数に絞り込むため、t-検定を行います。そのためにデータの変換ステップがいくつか必要になります。

Experimentの作成



- クラスタリングでは、発現量にRPKMを利用するため、再度Experiment setを作成します(発現差解析では、リード数を使ってください)。

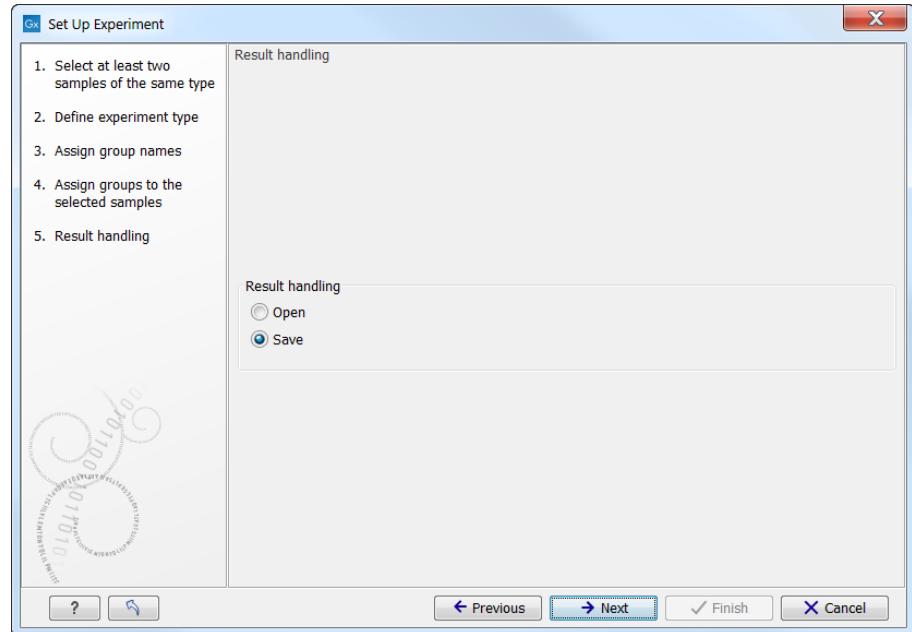
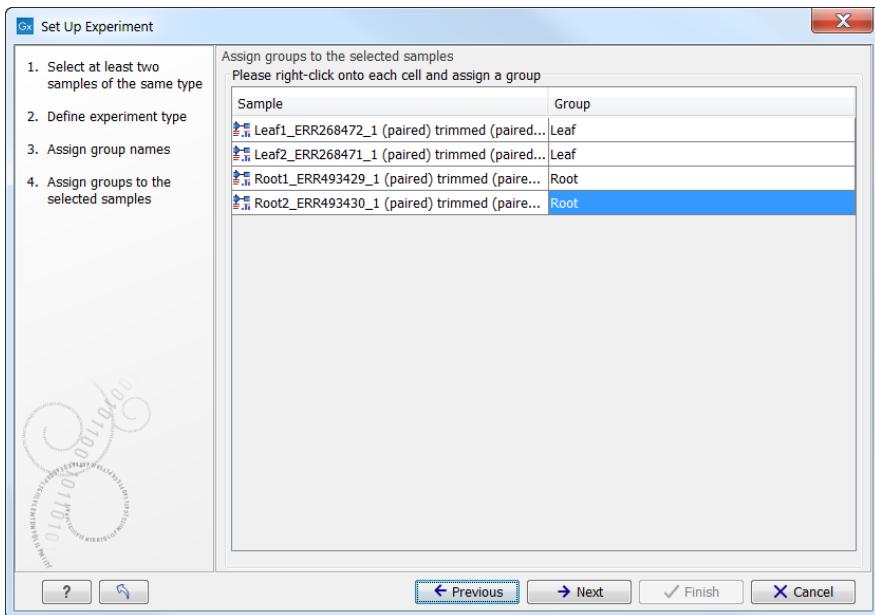
Experimentの作成



- Set new expression value でRPKMを選びます。

- 群の名前を設定します。

Experimentの作成

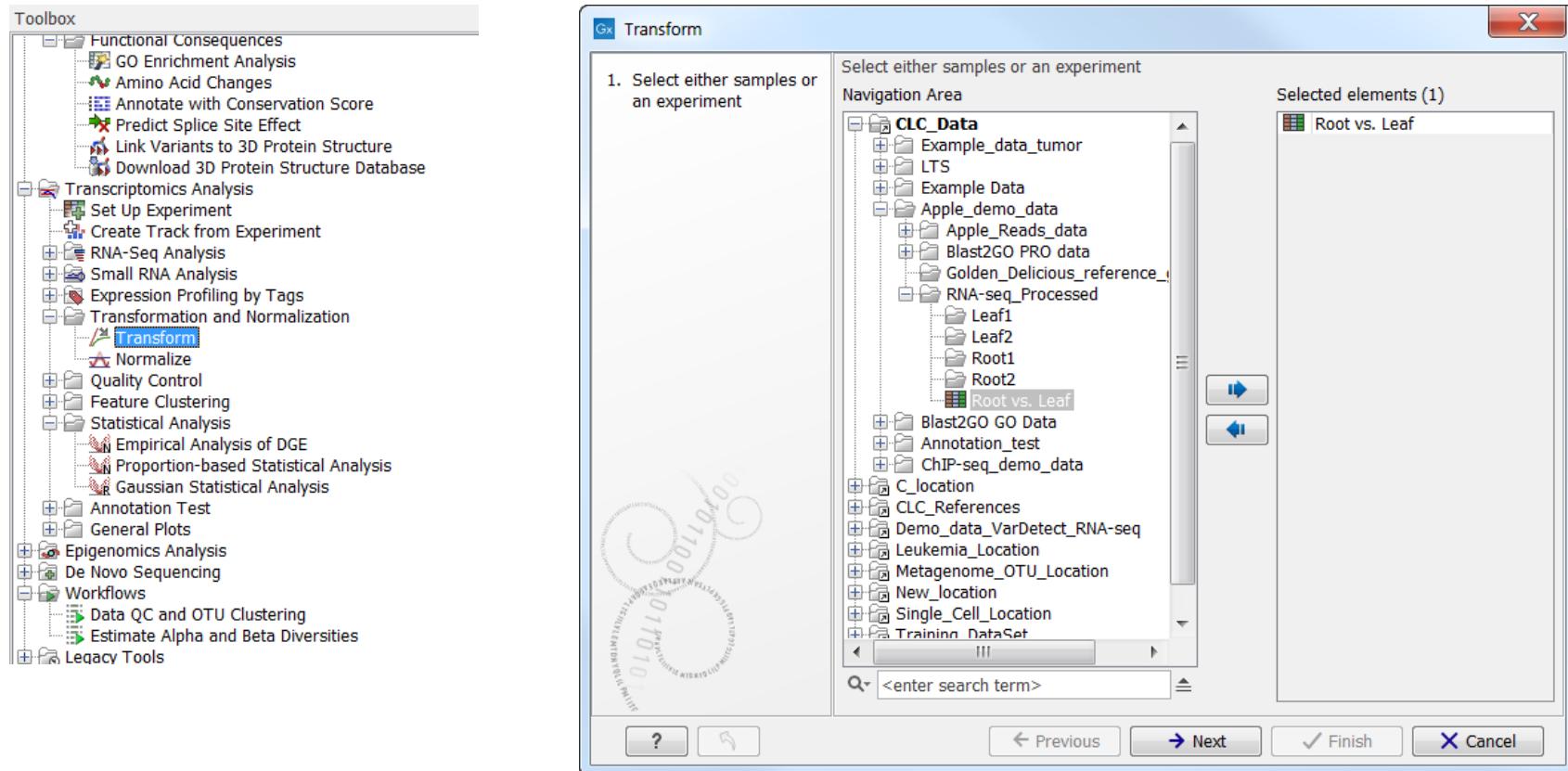


- 右クリックでRootとLeafに割り付けてます。

- 結果は下図のようになります。

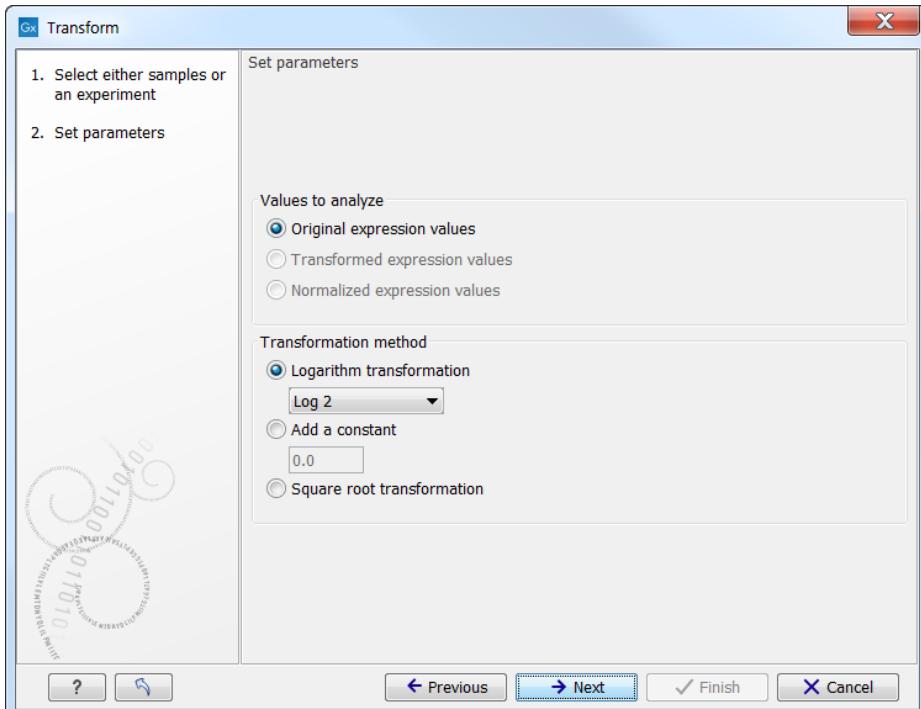
... Root vs. Leaf

ログ変換

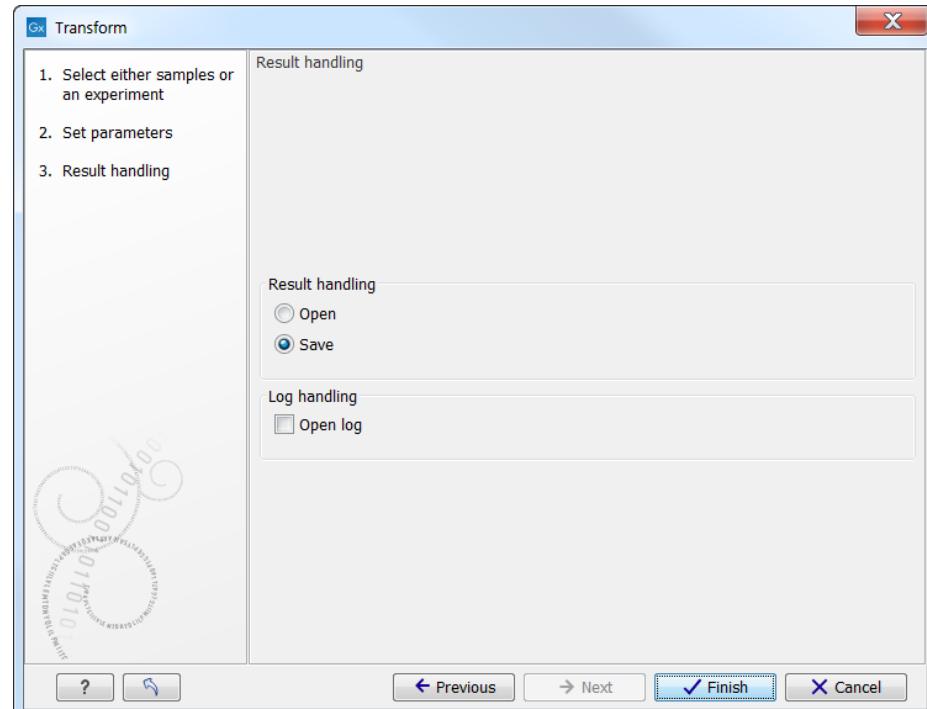


- Toolboxから NGS Core Tools > Transcriptomics Analysis > Transformation and Normalization > Transform を選択、ダブルクリック。
- 作成したExperimentを選択、保存します。

ログ変換



- ログ変換の底を選択



- 保存します。

結果

Root vs. Leaf

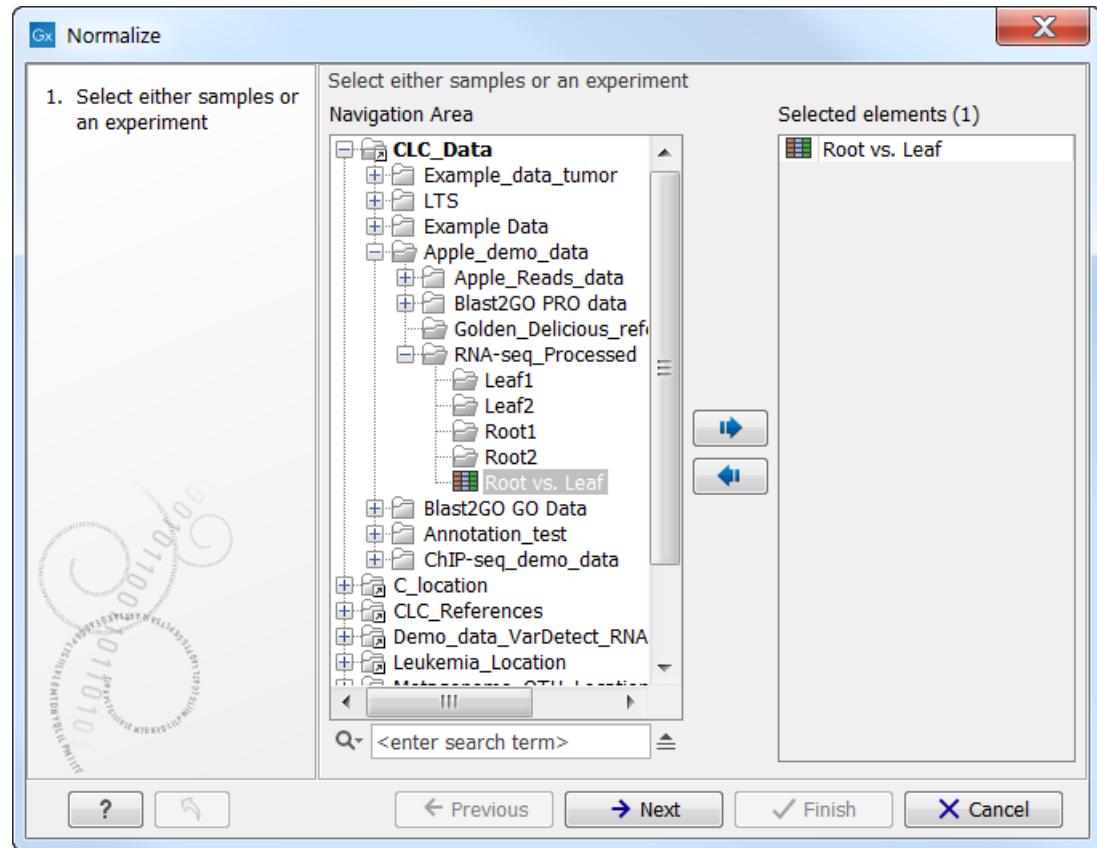
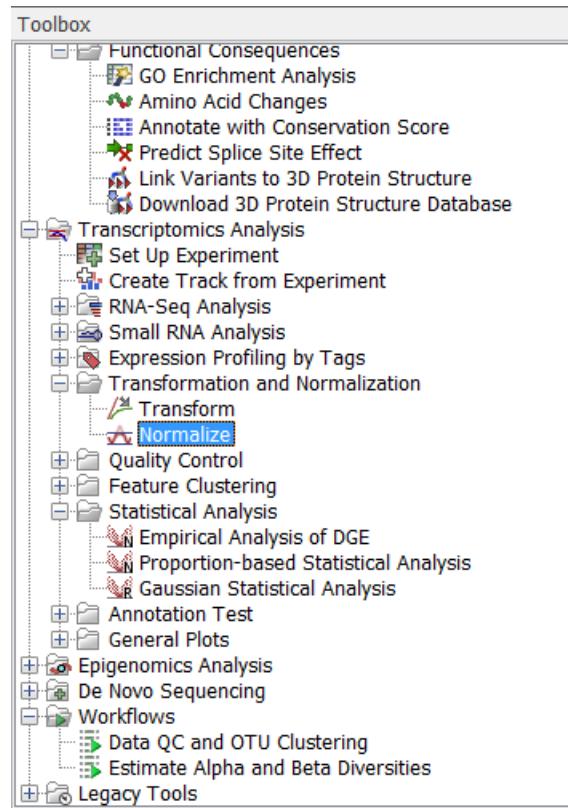
Chr16_NC24254... X Root vs. Leaf X

Rows: 1,415 Filter ▾

Difference ...	Fold Chan...			Root1_ERR493429_1 (paired) trimm								
		Expression values	Transformed values	Transcript...	Detected t...	Exon length	Unique ge...	Total gene...	Unique ex...	Total exon...	Ratio of un...	
-0.03	-1.00	136.50	7.09	3	1	1392	107	107	89	89	1.00	
-0.04	-1.00	261.95	8.03	1	1	4833	596	596	593	593	1.00	
-0.15	-1.02	269.80	8.08	1	1	1361	177	177	172	172	1.00	
-0.04	-1.01	389.34	8.60	2	1	3131	590	590	571	571	1.00	
-0.06	-1.01	395.72	8.63	1	1	1748	328	328	324	324	1.00	
-0.08	-1.01	1,769.42	10.79	1	1	3458	2884	2884	2866	2866	1.00	
-0.09	-1.01	1,147.05	10.16	1	1	1368	746	746	735	735	1.00	
-0.04	-1.01	81.47	6.35	1	1	891	34	34	34	34	1.00	
0.04	1.01	84.63	6.40	1	1	782	32	32	31	31	1.00	
0.01	1.00	120.37	6.91	1	1	3707	215	215	209	209	1.00	
-0.14	-1.02	266.68	8.06	1	1	1481	221	221	185	185	1.00	
0.05	1.01	463.24	8.86	1	1	2143	324	468	321	465	0.69	
-0.05	-1.01	197.56	7.63	1	1	1059	106	106	98	98	1.00	
0.14	1.01	847.61	9.73	1	1	2083	837	837	827	827	1.00	
0.02	1.00	293.77	8.20	2	2	1795	255	255	247	247	1.00	
-0.04	-1.00	146.32	7.19	1	1	1605	111	111	110	110	1.00	
0.21	1.04	29.60	4.89	1	1	1154	23	23	16	16	1.00	
-0.11	-1.01	769.81	9.59	2	2	2202	812	812	794	794	1.00	
0.15	1.02	143.29	7.16	1	1	2518	175	175	169	169	1.00	

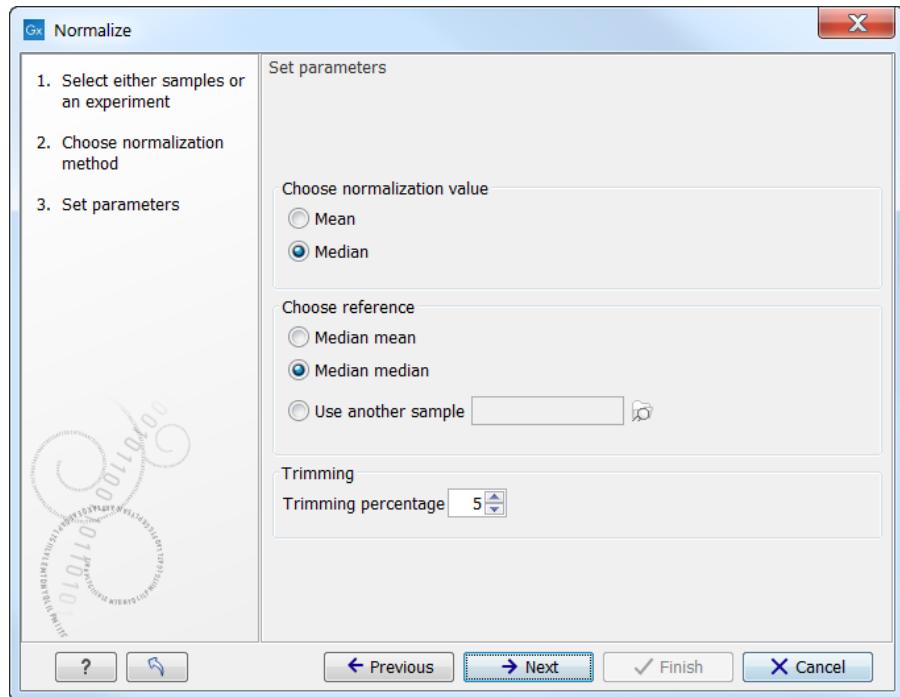
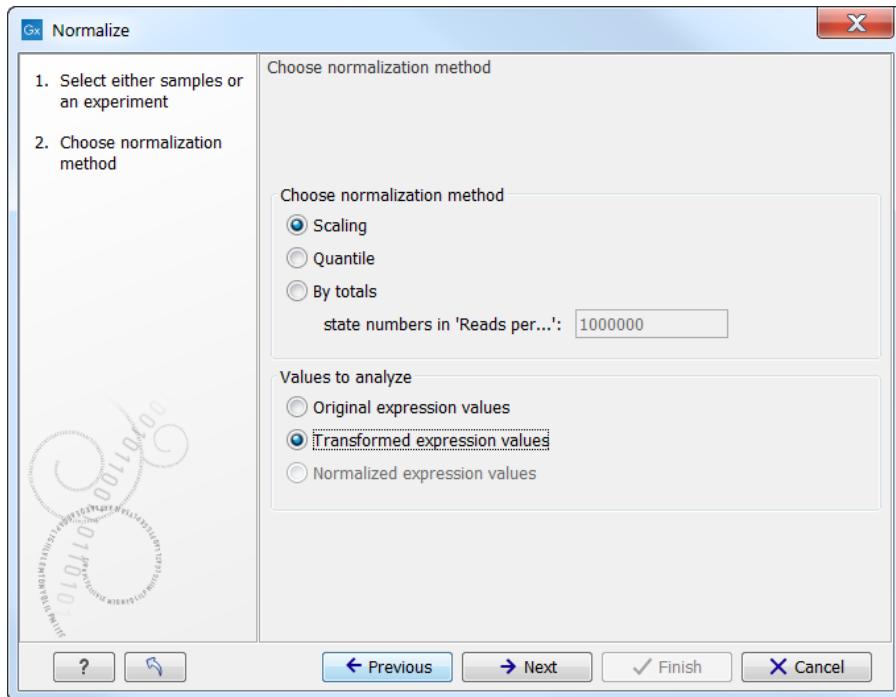
- ログ変換されたデータはExperiment Set に上書きされます。

ノーマライズ



- Toolboxから NGS Core Tools > Transcriptomics Analysis > Transformation and Normalization > Normalize を選択、ダブルクリック。
- 作成したExperimentを選択、保存します。

ノーマライズ



- ノーマライズ方法を選択。
- ノーマライズする値は、Transform後(ここではログ変換後)を選びます。

- ノーマライズに利用する値を選択。ここでは、前の画面で、Scalingを選択しているため、Scalingに使う方法を選択。
- Choose Reference では、ノーマライズに使う値を選択。
- 上下の値を一部トリミングすることができるのと、そのパーセンテージを選択

ノーマライズ

Gx Normalize

1. Select either samples or an experiment
 2. Choose normalization method
 3. Set parameters
 4. Result handling

Result handling
 Open
 Save

Log handling
 Open log

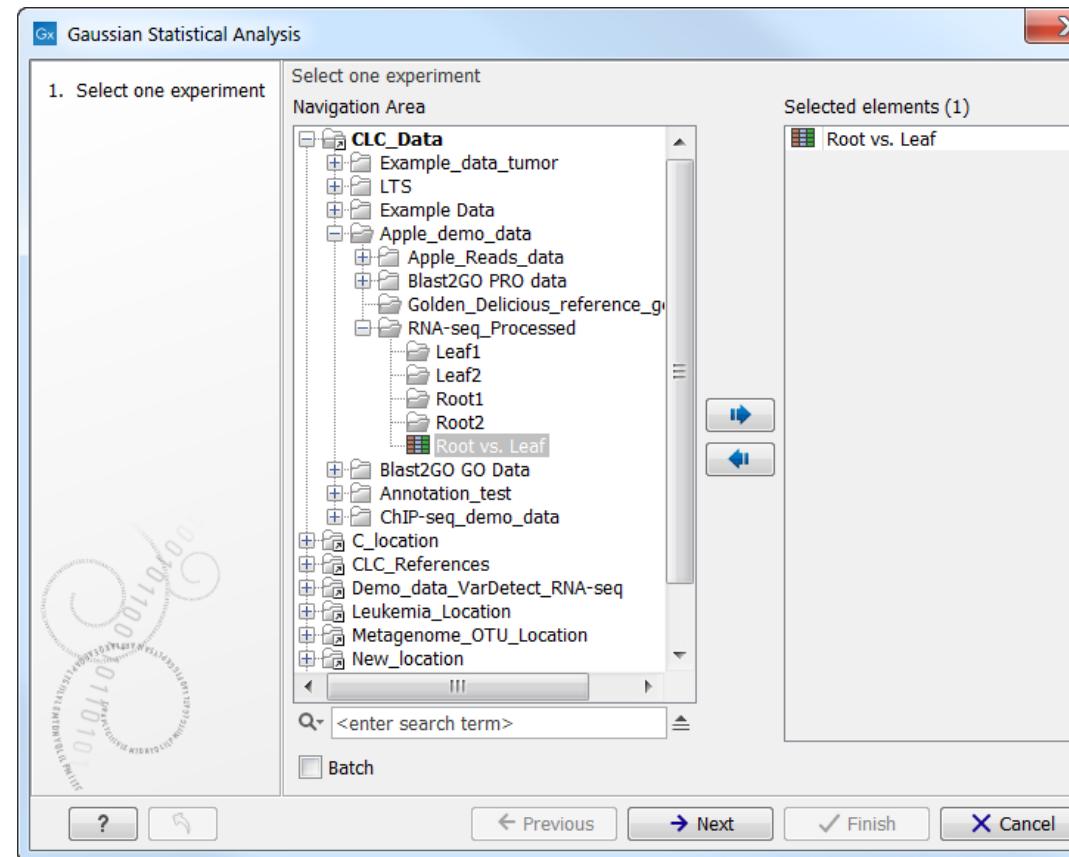
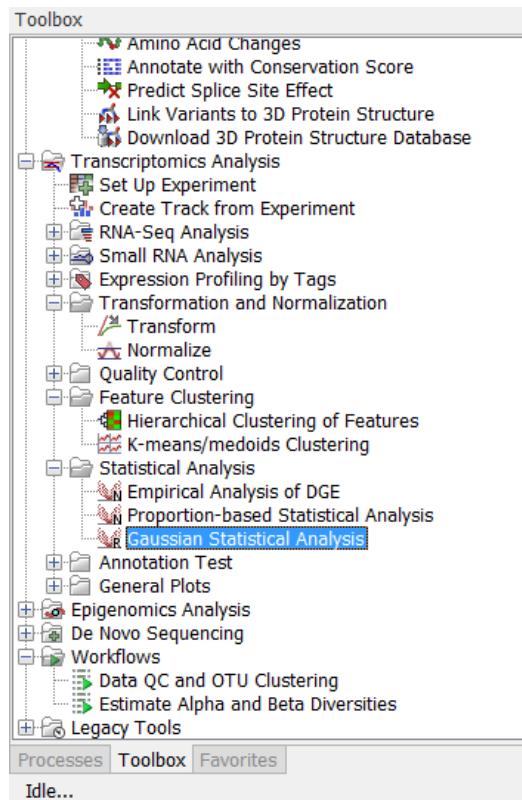
Root vs. Leaf

Rows: 1,415											
Difference ...	Fold Chan...	Range (no...)	IQR (norm...)	Difference ...	Fold Chan...	Expression values	Normalized express...	Transformed values	Transcript...	Detected t...	Exon le...
3.01	1.47	2.90	2.80	2.80	1.43	84.42	6.60	6.40	3	0	
0.21	1.02	0.50	0.40	0.05	1.01	618.04	9.50	9.27	1	1	
-9.74	-3.71	12.20	8.00	-9.95	-3.76	8,452.50	13.40	13.05	1	1	
-0.93	-1.11	1.60	0.80	-1.15	-1.14	660.05	9.60	9.37	1	1	
8.07	4.67	9.50	7.30	8.00	4.64	2.15	1.10	1.10	1	1	
-3.84	-1.45	4.80	3.50	-4.05	-1.48	5,124.32	12.60	12.32	1	1	
0.69	1.10	0.90	0.40	0.55	1.08	115.53	7.00	6.85	5	3	
0.75	1.11	1.60	1.00	0.55	1.08	89.64	6.70	6.49	1	1	
-2.88	-1.76	3.60	2.40	-3.00	-1.79	96.92	6.80	6.60	1	1	
0.17	1.02	1.10	0.90	-0.05	-1.01	376.19	8.80	8.56	1	1	
NaN	NaN	NaN	0.00	NaN	NaN	0.00	-∞	-∞	0	0	
2.19	1.45	3.20	2.70	2.05	1.41	23.77	4.70	4.57	1	1	
-1.35	-1.65	2.60	0.70	-1.40	-1.68	8.88	3.20	3.15	1	1	

- ノーマライズされた値は、Experimentに上書きされます。

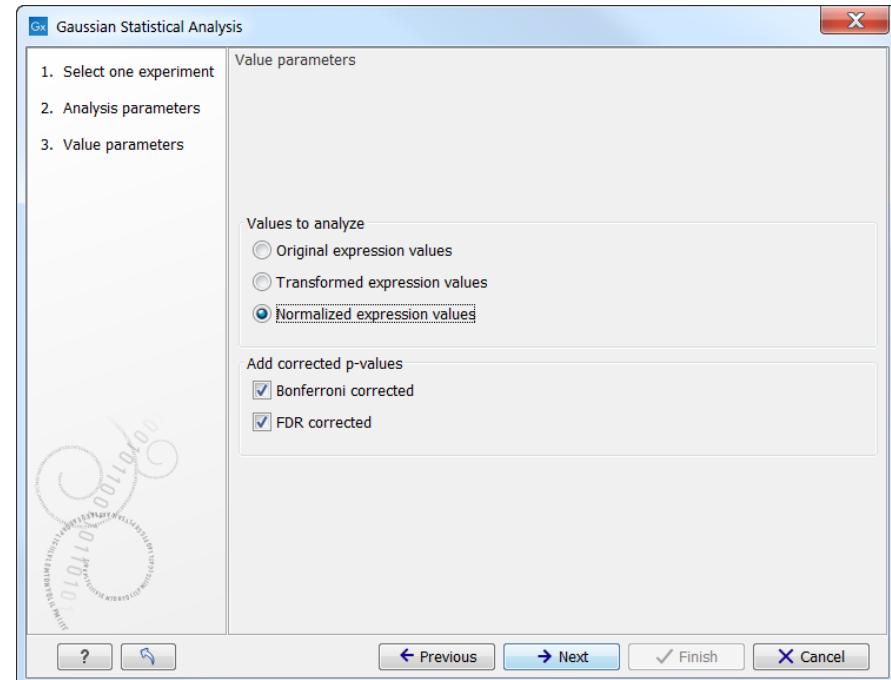
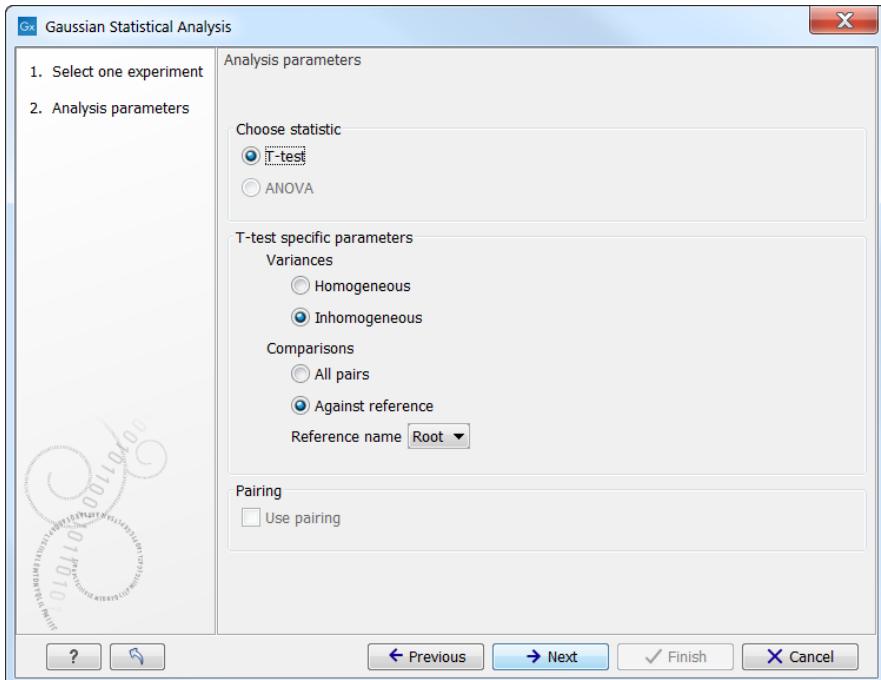
クラスタリング

t-検定



- Toolboxから NGS Core Tools > Transcriptomics Analysis > Gaussian Statistical Analysisを選択、ダブルクリック。
- 作成したExperimentを選択します。

t検定



- 2群の検定はt検定を、3群以上はANOVAを選択します。
- 分散の不均一性を選択し、対象群を選択します。

- 検定の値はノーマライズ後の値を選択します。
- p値の補正方法を選択します。

ノーマライズ

Rows: 1,415

Difference ...	Fold Chan...	Range (no...)	IQR (norm...)	Difference ...	Fold Chan...	Expression values	Normalized express...	Transformed values	Transcript...	Detected t...	Exon le...
3.01	1.47	2.90	2.80	2.80	1.43	84.42	6.60	6.40		1	1
0.21	1.02	0.50	0.40	0.05	1.01	618.04	9.50	9.27			
-0.74	-3.71	12.20	8.00	-9.95	-3.76	8,452.50	13.40	13.05			
-0.93	-1.11	1.60	0.80	-1.15	-1.14	660.05	9.60	9.37			
8.07	4.67	9.50	7.30	8.00	4.64	2.15	1.10	1.10			

• 黒の三角をクリック

Rows: 133 / 1,415

Feature ID	Range (original values)	IQR (original values)	Difference (original values)	Fold Chan...	Range (tra...)	IQR (trans...	Difference ...	Fold Chan...	Range (no...)
LOC103402583	619.03	617.41	603.53	8.08	3.06	3.03			
LOC103402598	113.46	113.11	100.97	5.13	2.50	2.48			
LOC103402608	991.28	965.73	-869.58	-18.41	4.79	4.04			
LOC103402617	673.83	660.89	583.79	2.78	1.63	1.57			
LOC103402630	97.04	84.74	85.83	2.76	1.71	1.34			
LOC103402633	197.40	196.55	-166.91	-112.57	7.53	6.69	-6.85	-14.17	7.70

• フィルターを変更設定

- t検定の結果をつかってフィルタリングします。デモデータでは補正したp値では23個の遺伝子のみが有意差をしめすのみとなるため、ここでは補正前のp値を使っていますが、本来の解析ではFDRまたはボンフェローニの値でフィルタリングしてください。
- フィルタリングで、ここでは2倍差のあるデータを選択(ログ変換しているので、2倍差を見るには、Different (Group2 – Group1)を選びます。)

クラスタリング

フィルタリング

Rows: 133 / 1,415

t-test: Root vs Leaf normalized values - P-value

Match any Match all

Filter

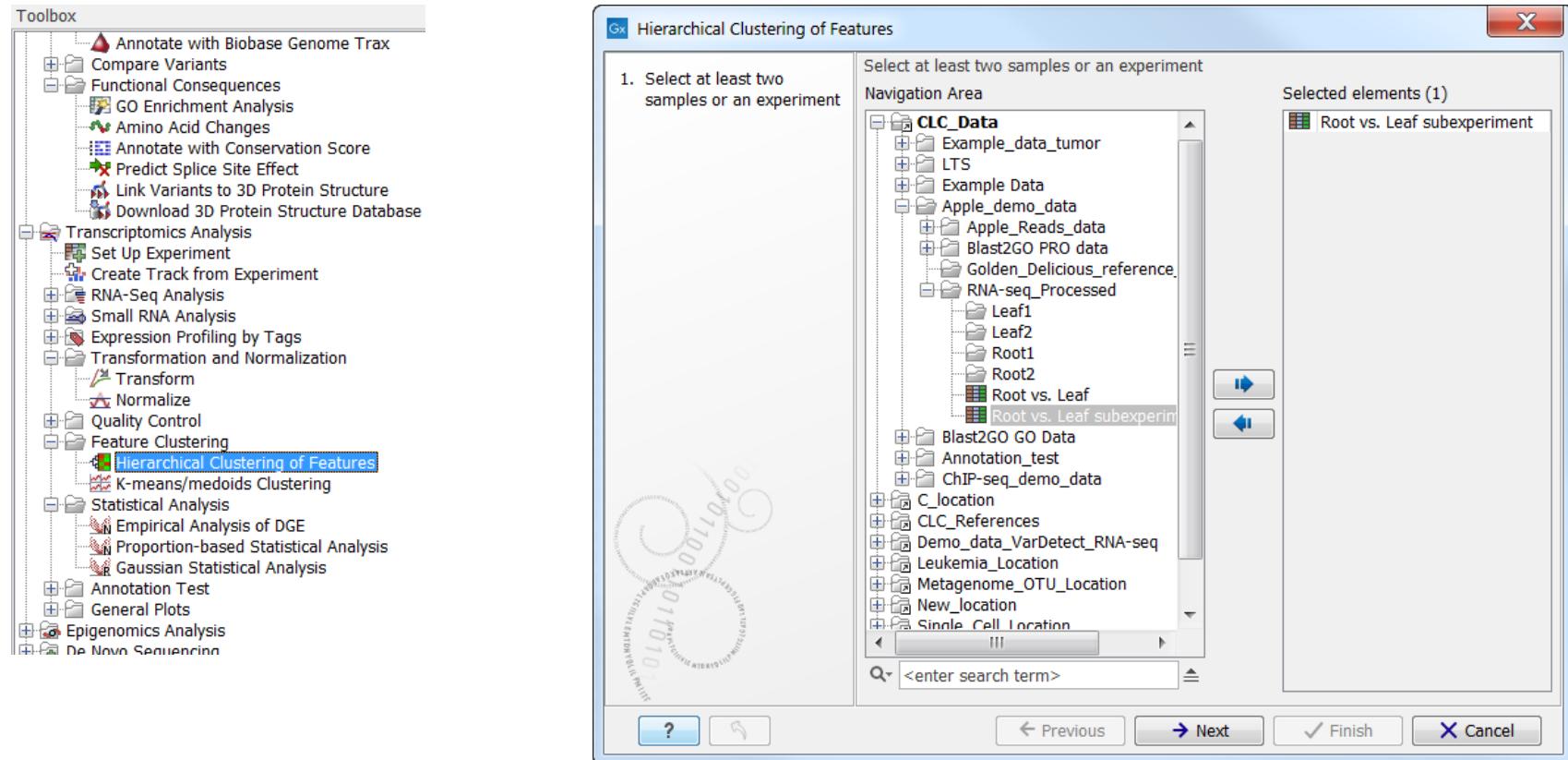
Feature ID	Experiment								
	Range (original values)	IQR (original values)	Difference (original values)	Fold Chan...	Range (tra...	IQR (trans...	Difference ...	Fold Chan...	Range (no...
LOC103402583	619.03	617.41	603.53	8.08	3.06	3.03	3.01	1.47	2.90
LOC103402598	113.46	113.11	100.97	5.13	2.50	2.48	2.35	1.51	2.50
LOC103402608	991.28	965.73	-869.58	-18.41	4.79	4.04	-4.24	-1.76	5.10
LOC103402617	673.83	660.89	583.79	2.78	1.63	1.57	1.47	1.18	1.50
LOC103402630	97.04	84.74	85.83	2.76	1.71	1.34	1.47	1.26	1.50
LOC103402633	197.40	196.55	-166.91	-112.57	7.53	6.69	-6.85	-14.17	7.70
LOC103402644	3,100.69	2,964.74	-2,787.37	-5.76	2.80	2.47	-2.53	-1.28	3.10
LOC103402653	172.54	168.70	-156.45	-3.59	1.98	1.89	-1.84	-1.31	2.10
LOC103402654	21.71	20.52	17.90	3.82	2.26	1.98	1.93	1.72	2.10
LOC103402663	258.13	249.42	-217.61	-7.23	3.24	2.88	-2.85	-1.56	3.50
LOC103402675	165.51	167.78	-159.38	-24.84	4.76	4.61	-4.64	-2.69	5.00
LOC103402686	58.90	57.13	50.61	6.39	2.99	2.71	2.67	1.83	2.90
LOC103402698	4,003.65	3,977.26	-3,342.51	11.27	3.79	3.67	3.47	1.42	3.60
LOC103402699	27.40	24.67	20.93	3.45	2.27	1.80	1.78	1.58	2.20
LOC103402711	1,304.58	1,201.29	-1,113.15	-4.93	2.73	2.20	-2.32	-1.29	3.10
LOC103402716	560.67	491.61	475.23	2.97	1.90	1.48	1.58	1.20	1.60
LOC103402725	320.13	302.98	250.17	5.90	3.09	2.60	2.55	1.45	2.90
LOC103402746	354.50	343.71	-340.40	-13.53	4.11	3.53	-3.79	-1.80	4.40
LOC103402784	94.92	94.32	-89.79	-17.09	4.25	4.09	-4.10	-2.65	4.40
LOC103402790	676.12	648.40	654.81	8.20	3.29	2.84	3.05	1.47	3.10
LOC103402801	2,565.53	2,509.55	-2,293.79	25.94	5.36	4.45	4.76	1.74	5.10
LOC103402808	145.95	127.88	-110.03	-2.28	1.53	1.23	-1.18	-1.18	1.60
LOC103402809	20.81	20.19	17.84	10.20	3.78	3.32	3.36	4.58	3.80
LOC103402825	1,275.02	1,040.77	-1,010.70	-2.26	1.52	1.09	-1.19	-1.12	1.80
LOC103402842	796.57	678.25	-669.46	-2.51	1.62	1.23	-1.34	-1.15	2.00
LOC103402843	1,978.21	1,973.71	1,765.82	227.55	8.48	7.63	7.88	3.72	8.40
LOC103402854	5,782.37	5,765.29	5,298.50	3.73	2.00	1.99	1.90	1.17	1.80
LOC103402861	1,520.14	1,319.99	-1,410.52	-2.00	1.11	0.91	-1.00	-1.10	1.50
LOC103402863	850.78	833.50	817.74	9.32	3.39	3.14	3.23	1.49	3.30
LOC103402870	272.12	262.99	227.42	12.05	4.17	3.52	3.61	1.83	4.00
LOC103402871	521.86	403.40	400.06	1.98	1.32	0.90	1.00	1.12	1.00
LOC103402898	12,186.58	12,045.78	-11,144.36	-11.12	3.68	3.50	-3.47	-1.34	4.10
LOC103402902	454.54	435.67	407.19	5.06	2.58	2.31	2.34	1.35	2.30
LOC103402915	265.31	249.95	237.10	2.67	1.57	1.42	1.42	1.20	1.40
LOC103402917	1,563.95	1,478.67	-1,321.55	-6.19	3.07	2.58	-2.64	-1.33	3.40
LOC103402924	5,410.82	5,107.87	-5,182.69	-3.56	1.96	1.74	-1.84	-1.17	2.30
LOC103402939	58.46	54.44	-46.03	-1.85	1.08	0.97	-0.88	-1.15	1.10
LOC103402951	275.84	274.97	-235.72	-81.46	6.80	6.37	-6.34	-5.13	6.90
LOC103402958	837.94	837.89	788.40	564.02	9.25	9.20	9.14	19.82	9.20
LOC103402973	2,217.14	1,936.50	-1,817.76	-3.10	2.02	1.55	-1.64	-1.17	2.30
LOC103402989	798.98	798.24	-756.09	-2.47	1.35	1.35	-1.31	-1.14	1.60
LOC103403002	15,210.39	15,208.78	10,066.59	1,997.94	11.81	11.35	10.77	5.65	11.70
LOC103403013	59.65	58.59	56.46	3.89	2.05	1.97	1.96	1.46	2.00
LOC103403025	39.71	34.70	-35.44	-1.26	0.38	0.32	-0.33	-1.05	0.60
LOC103403030	2,209.60	1,908.79	-1,800.05	-3.28	2.15	1.60	-1.73	-1.18	2.50
LOC103403056	80.58	80.22	-75.56	-10.79	3.55	3.48	-3.43	-2.16	3.70
LOC103403064	1,512.72	1,511.27	1,173.65	258.84	8.63	8.17	7.97	4.68	8.60
LOC103403069	2,257.98	2,256.82	1,857.60	224.23	8.19	7.99	7.78	3.55	8.10
LOC103403070	12,551.63	12,170.51	-11,618.98	-10.18	3.66	3.23	-3.36	-1.33	4.10

Add Annotations Create Experiment from Selection Download Sequence

Root vs. Leaf subexperiment

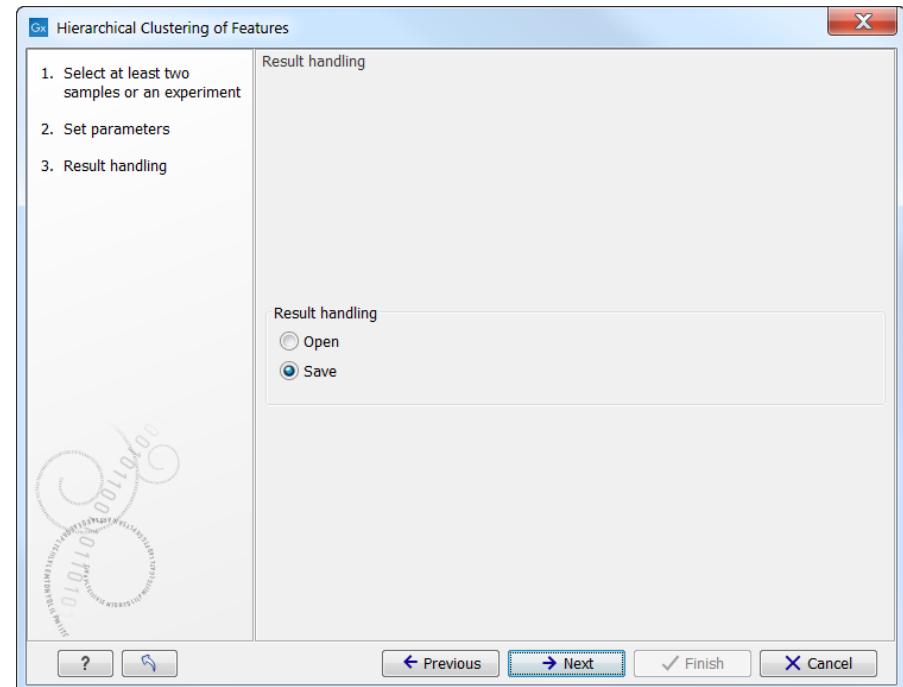
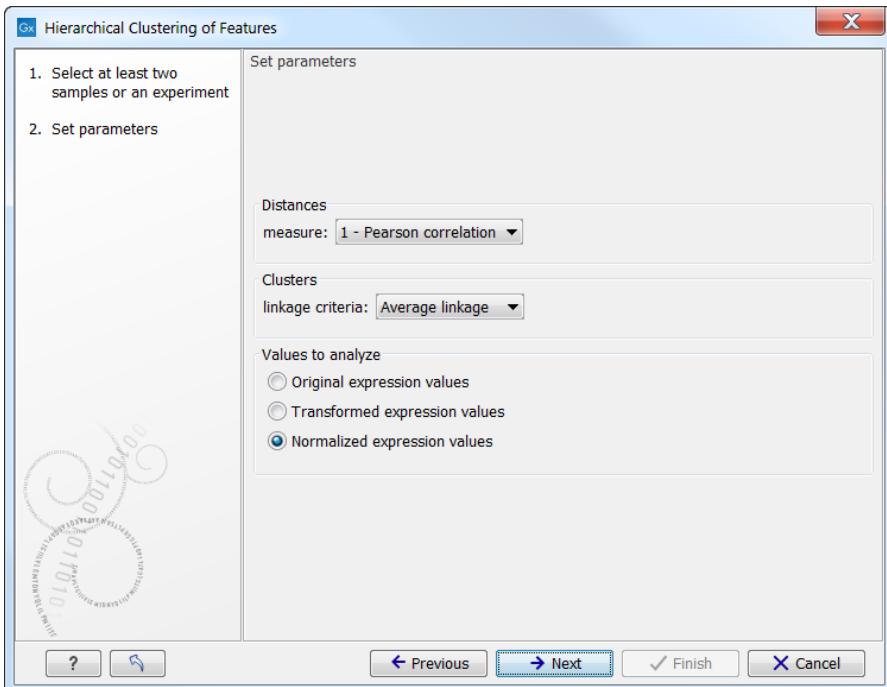
- フィルタリングされた値をCtrl + Aで全選択し、Create Experiment from Selection を選択します。
- これにより、選択した値のみのExperiment Set が作成されます。これを使って、クラスタリングを行います。

クラスタリング



- Toolbox > Transcriptomics Aanlysis > Feature Clustering > Hierarchical Clustering of Features を選択。
- 作成したサブセットのExperimentが選択されていることを確認。

クラスタリング

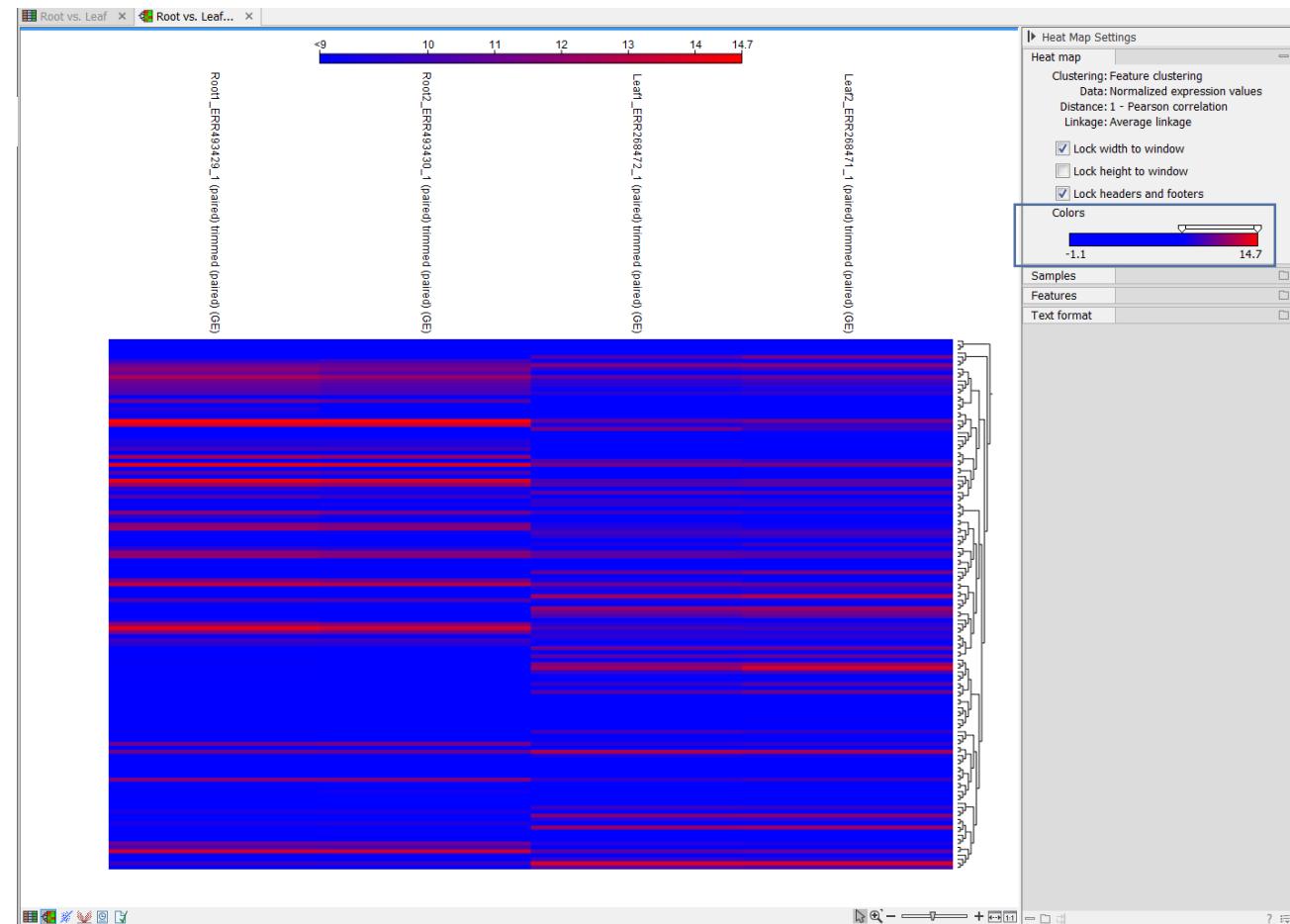


- Distances ではクラスタリングの類似度に使う値を選択
- クラスタリングの代表値を選択(階層的クラスタリングでは、次のクラスタリングに使用)
- クラスタリングに使う値を選択。

- 結果の保存

結果

Root vs. Leaf subexperiment

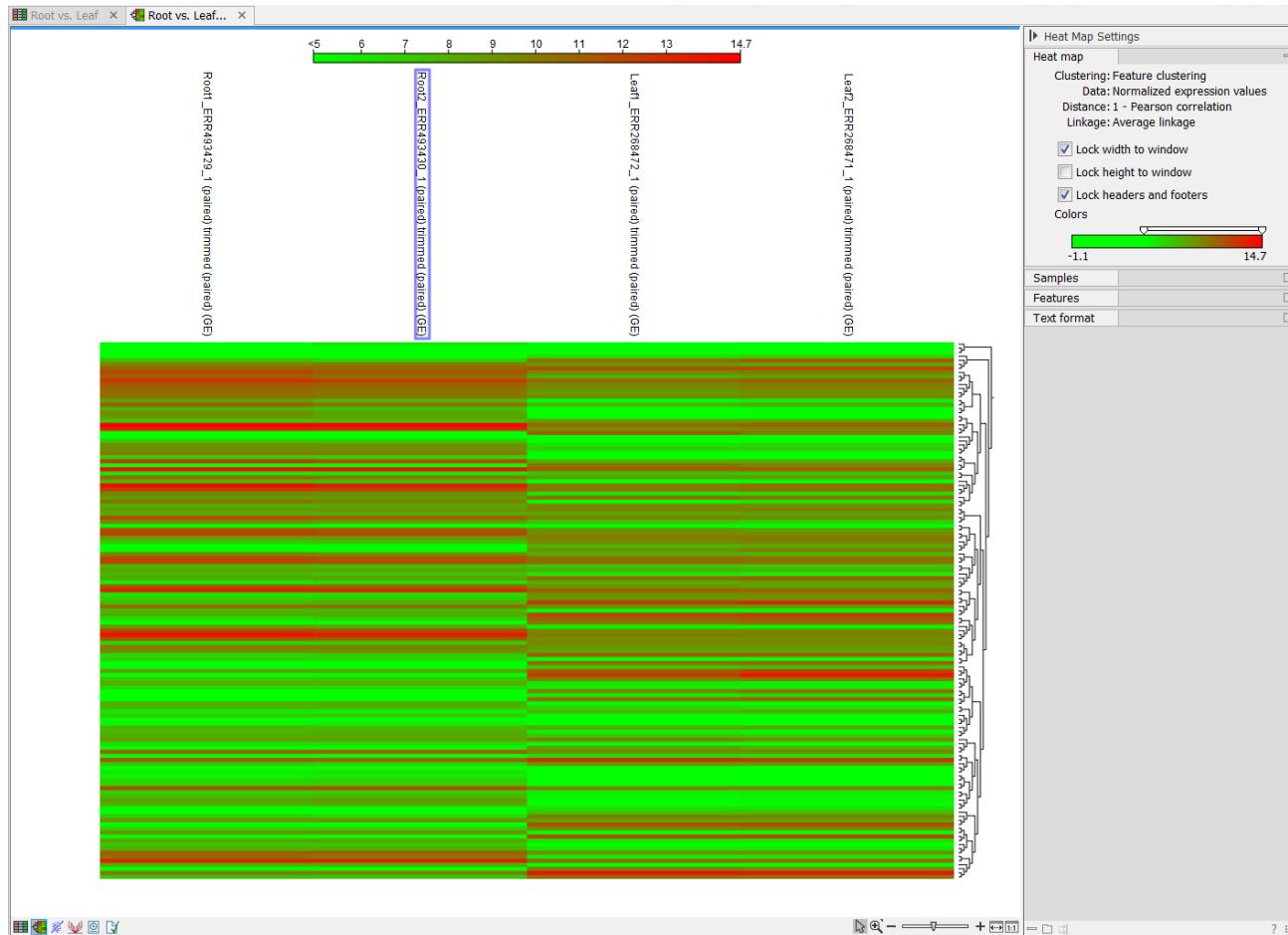


- 結果はExperiment Set のビューを切り替えることで表示。
- 右側のカラーバーを変更し、好きなように色と濃淡を変更。



クラスタリング

結果



- このような表示にすることも可能。

Expression analysis – 補足 : P値の補正

- 検定を繰り返すと、指定した閾値よりも実際は高いエラーを含むことになります。たとえば、 $p < 0.05$ となる遺伝子のリストを得たい場合、3つの遺伝子について検定を行った場合、これは検定の繰り返しとなり、実際には $1 - (1 - 0.05)^3 = 0.14$ というエラーを含んだ結果となるのです。
- Bonferroni 法ではくりかえしの検定数で閾値を割ることで、繰り返しを考慮した閾値を設定します。上記の例では、 $0.05 / 3$ とした閾値で検定します。
- しかし、遺伝子数が膨大になると、閾値が非常に小さくなり、どの遺伝子も検定で棄却できず、リストが作成できなくなり、現実的ではありません。
- ボンフェローに法の閾値で棄却されたリストと言うのは、False Positiveを全く含まないリストとなります。これを少し緩くし、ある程度のエラーを含むことを覚悟した上でリストを得ようとする方法がFDRになります。

Expression analysis – 補足 : RPKM

$$\text{RPKM} = \frac{\text{total exon reads}}{\text{mapped reads(millions)} \times \text{exon length (KB)}}$$

- Total exon reads: exon にマップされたリード数
- Exon length: 遺伝子のすべての exon を足して 1000 で割った数
- Exon が長いほどマップされるリードが多くなるので、exon length で割ることにより補正する
- Mapped reads: マップされたリードの総数を 100万で割った数

お疲れ様でした。

