

RNAseq report manual

1. 用語説明.....	3
2. Html レポートの全体像	3
3. 解析内容.....	4
3.1. 解析サンプル情報	4
3.2. 変動遺伝子解析パターン.....	4
3.3. 解析結果概要	4
3.3.1. リード品質	4
3.3.2. 遺伝子数	4
4. 品質管理.....	5
4.1. General statistics	5
4.2. FastQC.....	5
4.2.1. Sequence Counts.....	5
4.2.2. Sequence Quality Histograms	5
4.2.3. Per Sequence Quality Scores	5
4.2.4. Per Sequence GC Content	5
4.2.5. Per Base N Content	5
4.2.6. Adapter Content	5
4.2.7. Status Checks	6
4.3. fastp.....	6
4.3.1. Filtered Reads	6
4.3.2. Insert Sizes.....	6
4.3.3. Sequence Quality.....	6
4.3.4. GC content	6
4.3.5. N content	6
4.4. HiSAT2	6
4.5. featureCounts	6
5. 遺伝子発現解析	8
5.1. リードカウントデータ.....	8
5.2. 遺伝子数	8
5.3. 遺伝子リード数	8
5.4. 主成分分析	8
5.5. 相関分析	8
5.6. クラスタリング	8
6. 変動遺伝子解析	9
6.1. 変動遺伝子解析結果	9

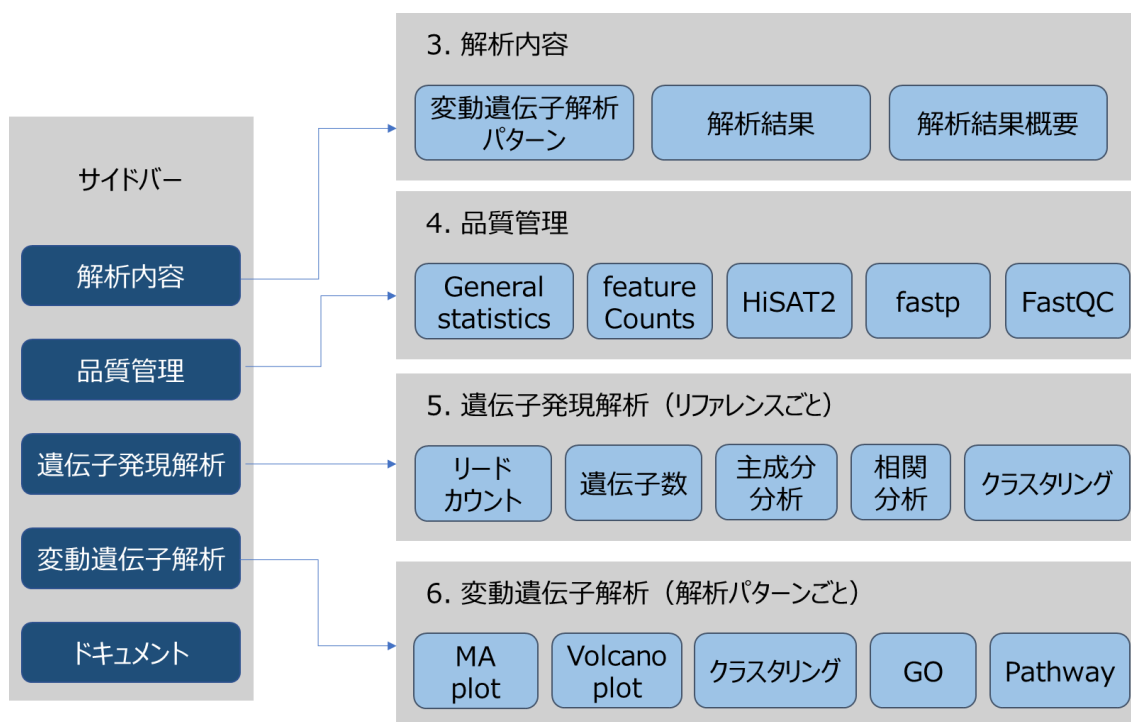
6.2. MA plot.....	9
6.3. Volcano plot.....	9
6.4. クラスターリング解析.....	9
6.5. GO	9
6.5.1. Bar plot.....	10
6.5.2. Category Netplot	10
6.5.3. Hierarchical plot.....	10
6.6. Pathway.....	10
6.6.1. Barplot.....	11
6.6.2. Category Netplot	11
7. ドキュメント.....	12

1. 用語説明

Phred Score	リードの品質を表す数値です。 $-10 \times \log_{10}(\text{エラー率})$ で算出されます。例えば、Phred Score が 30 の場合、エラー率は 0.1% で読み取られた塩基の信頼度は 99.9% となります。
FPKM	fragments per kilobase of exon per million reads mapped の略。総リード数を 1,000,000 リードに補正した後、転写産物 1,000 bp あたりのリード数を算出する方法。
TPM	transcripts per million の略。転写産物 1,000 bp あたりのリード数に補正した後、総リードカウントが 1,000,000 となるように補正する方法。
GO	Gene Ontology の略。生物学的領域に関する情報を Biological process (BP)、Cellular component (CC)、Molecular function (MF) 3つの側面から記述し、整理したもの。
Reactome	ヒトなどの生体内での反応やパスウェイを整備したデータベース。

2. Html レポートの全体像

解析レポートの全体像です。図中の番号は本マニュアルの章番号です。



3. 解析内容

3.1. 解析サンプル情報

RNAseq 解析を行ったサンプルについての表です。

3.2. 変動遺伝子解析パターン

変動遺伝子解析のグループについてまとめた表です。

3.3. 解析結果概要

3.3.1. リード品質

fastp による低品質リードの除去結果です。

項目名	説明
Passed Filter	低品質リードが除去され、高品質塩基のみとなったリード数です。
Low Quality	低品質塩基を除去後、リードの平均 Phred Score が 15 未満となったリード数です。
Too Many N	リード内に 6 塩基以上の未決定塩基（N）が含まれていたリード数です。
Too Short	低品質塩基除去後、40 塩基未満になったリード数です。

3.3.2. 遺伝子数

少なくとも 1 以上のリードがマッピングされた遺伝子数の棒グラフです。

項目名	説明
protein_coding	タンパク質をコードする遺伝子です。
rRNA	Ribosome RNA です。
ncRNA	rRNA 以外のタンパク質をコードしない遺伝子/転写産物です。
pseudogene	偽遺伝子です。

4. 品質管理

この項目ではリードの品質やリファレンス配列へのマッピング結果についてまとめています。

4.1. General statistics

列名	説明
Sample	サンプル名です。
M Seqs	全リード数です（単位：100 万リード）。
Read Length	リードの長さです。
% Adapter	アダプターが含まれていたリードの割合です。
% GC content	リードの GC 含有率です。
M PF	高品質リード数です（単位：100 万リード）。
% PF	全リードの内、高品質リードの割合です。
M Aligned	リファレンス配列にマッピングされたリード数です（単位：100 万リード）。
% Aligned	高品質リードの内、リファレンス配列にマッピングされたリードの割合です。
M Assigned	リファレンス配列にマッピングされたリードの内、遺伝子にマップされたリード数です（単位：100 万リード）。
% Assigned	リファレンス配列にマッピングされたリードの内、遺伝子にマップされた割合です。

4.2. FastQC

4.2.1. Sequence Counts

FastQC により算出されたユニークなリード数と重複したリード数です。横軸の単位は M (10^6) read です。

4.2.2. Sequence Quality Histograms

リード内のリードクオリティ分布です。横軸はリード内における塩基の位置で、縦軸はリード位置ごとの平均 Phred score です。

リードの位置における品質のばらつきを確認出来ます。

4.2.3. Per Sequence Quality Scores

リードクオリティスコア（Phred score）のリード数ヒストグラムです。

リードの品質が安定しているかを確認出来ます。

4.2.4. Per Sequence GC Content

GC 含有率のリード数ヒストグラムです。

4.2.5. Per Base N Content

未決定塩基（N）含有率のリード数ヒストグラムです。

4.2.6. Adapter Content

「Position in Read」までにアダプターが検出されたリードの割合です。

アダプターの混入状況を確認出来ます。

4.2.7. Status Checks

FastQC による各種チェックの結果です。各項目の評価基準は FastQC のサイトに説明が記載されています。[\(評価方法\)](#)

Status は pass、warn、fail があり、それぞれ pass は問題なし、warn と fail は注意が必要な項目です。

4.3. fastp

4.3.1. Filtered Reads

fastp による低品質リードの除去結果です。詳しくは [3.3.1.](#)に記載しています。

4.3.2. Insert Sizes

fastp により算出されたリードペアのインサートサイズの分布です。

4.3.3. Sequence Quality

低品質リード除去前後のリード内リードクオリティ分布です。横軸がリード内の位置、縦軸が Phred score です。

4.3.4. GC content

低品質リード除去前後の GC 含有率です。横軸がリード内の位置、縦軸は GC 含有率です。

4.3.5. N content

低品質リード除去前後の未決定塩基（N）の含有率です。横軸がリード内の位置、縦軸は未決定塩基の割合です。

4.4. HiSAT2

リファレンス配列に対してリードマッピングした結果を棒グラフで表示しています。

項目名	説明
mapped uniquely	リファレンス配列に一箇所のみマッピングされたリード数です。
multimapped	リファレンス配列に複数回マッピングされたリード数です。
neither mate aligned	リードペアでマッピングされなかったリード数です。

4.5. featureCounts

マッピングされたリードについて、featureCounts により集計した結果を棒グラフで表示しています。

項目名	説明
Assigned	正しく遺伝子にマッピングされたリード数です。
Unassigned: Ambiguity	複数の遺伝子に重なるようにマッピングされたリード数です。

Unassigned: MultiMapping	リファレンス配列に複数回マッピングされたリード数です。
Unassigned: NoFeatures	非遺伝子領域にマッピングされた遺伝子数です。
Unassigned: others	その他の理由により正しくマッピングされなかったリード数です。

5. 遺伝子発現解析

5.1. リードカウントデータ

各種リードカウントデータは下記の構成となっています。

列名	説明
Gene ID、Geneid	遺伝子 ID です。
Gene Name、gene_name	遺伝子名です。
[Sample_ID]	サンプルごとのカウントデータです。

5.2. 遺伝子数

少なくとも 1 リード以上がマッピングされた遺伝子数の棒グラフです。詳しくは [3.3.2.](#)に記載しています。

5.3. 遺伝子リード数

遺伝子にマップされたリード数を横軸にした棒グラフです。

5.4. 主成分分析

PCA データの表は下記の構成となっています。

列名	説明
Sample_ID	サンプル名です。
Group_ID	グループ名です。
PCXX	主成分得点です。PC1 は第一主成分得点を表しています。

プロットの横軸は第一主成分、縦軸は第二主成分です。軸タイトルの%の数値は各主成分の寄与率を表しています。

5.5. 相関分析

相関分析結果の表は行、列ともにサンプル名で、相関係数のクラスティング順に並んでいます。表中の数値は相関係数です。

5.6. クラスティング

TPM > 10 の遺伝子を対象にクラスティングした結果をプロットしています。

クラスティング解析結果の表は下記の構成となっています。

遺伝子 ID と遺伝子名、サンプル名はクラスティング結果の順に並んでいます。

列名	説明
Gene ID	遺伝子 ID です。
Gene Name	遺伝子名です。
[Sample_ID]	サンプルごとの TPM 正規化データです。

6. 変動遺伝子解析

6.1. 変動遺伝子解析結果

変動遺伝子解析の結果です。

列名	説明
Geneid	遺伝子 ID です。
baseMean	正規化カウント数の平均値です。
log2FoldChange	変化量の対数です。
pvalue	検定結果の P 値です。
padj	BH 方による P 値の補正值です。
estimatedDEG	padj ≥ 0.05 かつ log2FoldChange の絶対値 ≥ 1 を満たす遺伝子が「1」、それ以外が「0」です。
gene_name	遺伝子名です。
gene_biotype	遺伝子の種類 (protein_coding、ncRNA、pseudogene) です。

6.2. MA plot

横軸が baseMean の log2 値、縦軸が Log2FoldChange のプロットです。

赤色は有意に増加した遺伝子、青色は減少した遺伝子、灰色は変動なしの遺伝子を表しています。

6.3. Volcano plot

横軸が Log2FoldChange、縦軸が $-\log_{10}(\text{padj})$ のプロットです。

赤色は、padj ≤ 0.05 , Log2FoldChange の絶対値 ≥ 1 を満たす有意に変動した遺伝子です。緑色は log2 Fold Change ≥ 1 のみを満たす遺伝子です。青色は Adjusted P value ≤ 0.05 のみを満たす遺伝子です。灰色は両方を満たしていない遺伝子です。

6.4. クラスタリング解析

変動した遺伝子を対象にクラスタリング結果をプロットしています。

表の構成は 5.4.と同じです。

6.5. GO

Upregulate と Downregulate した遺伝子それぞれで GO 解析した結果です。

GO 解析結果の評価下記の構成となっています。

列名	説明
ONTOLOGY	ONTOLOGY 名です。BP、CC、MF のいずれかが記載されています。
ID	GO ID です。
Description	GO の名前です。
GeneRatio	[GO ID に含まれる変動遺伝子数 (LH)] / [変動遺伝子数 (LT)] です。
BgRatio	[GO ID に含まれる全遺伝子数 (PH)] / [GO アノテーションを持つ全遺伝子数 (PT)] です。
pvalue	超幾何分布による検定結果の P 値です。

p.adjust	Benjamini-Hochberg 法による P 値の補正值です。
qvalue	qvalue パッケージによる P 値の補正值です。
geneID	DEG 中で GO ID に含まれる遺伝子名です。
Count	DEG 中で GO ID に含まれる遺伝子数です。

6.5.1. Bar plot

変動した遺伝子が多い GO term を棒グラフで表しています。

縦軸は有意にエンリッチされた GO term、横軸は $-\log_{10}(\text{padj})$ です。

棒グラフの色は下記で算出されるオッズ比です。

$$\text{Odds Ratio} = \frac{LH / (LT - LH)}{(PH - LH) / (PT - LT - (PH - LH))}$$

6.5.2. Category Netplot

エンリッチされた GO term と遺伝子のネットワーク図です。有意にエンリッチされた GO term とその GO term 内の有意に変動した遺伝子についてプロットしています。

このプロットによりエンリッチされた GO term 同士の関係性を確認することができます。

6.5.3. Hierarchical plot

有意にエンリッチされた GO term の上位階層の GO term を可視化したプロットです。

このプロットにより、エンリッチされた GO term の上位階層の GO term、エンリッチされた GO term 同士の関係性を確認することができます。

6.6. Pathway

Reactome データベース内のパスウェイのエンリッチメント解析結果です。

Pathway 解析結果の表は下記の構成となっています。

列名	説明
ID	Reactome Pathway ID です。
Description	Reactome Pathway の名前です。
GeneRatio	[Reactome Pathway ID に含まれる変動遺伝子数 (LH)] / [変動遺伝子数 (LT)] です。
BgRatio	[Reactome Pathway ID に含まれる全遺伝子数 (PH)] / [Reactome Pathway のアノテーションを持つ全遺伝子数 (PT)] です。
pvalue	超幾何分布による検定結果の P 値です。
p.adjust	Benjamini-Hochberg 法による P 値の補正值です。
qvalue	qvalue パッケージによる P 値の補正值です。

geneID	DEG 中で Reactome Pathway ID に含まれる遺伝子名です。
Count	DEG 中で Reactome Pathway ID に含まれる遺伝子数です。

6.6.1. Barplot

変動した遺伝子が多いパスウェイを棒グラフで表しています。

縦軸は有意にエンリッチされたパスウェイ、横軸は $-\log_{10}(\text{padj})$ です。

棒グラフの色は下記で算出されるオッズ比です。

$$\text{Odds Ratio} = \frac{LH / (LT - LH)}{(PH - LH) / (PT - LT - (PH - LH))}$$

6.6.2. Category Netplot

エンリッチされたパスウェイと遺伝子のネットワーク図です。有意にエンリッチされたパスウェイとそのパスウェイ内の有意に変動した遺伝子についてプロットしています。

このプロットによりエンリッチされたパスウェイ同士の関係性を確認することができます。

7. ドキュメント

下記の内容にアクセス出来ます。

- 解析手順書
RNAseq 解析の手順を確認出来ます。
- 納品ファイル説明
各種納品物についての説明を記載しているファイルにアクセス出来ます。
 1. gene_expression_analysis.txt: 遺伝子発現解析に関するファイルについての説明です。
 2. degs.txt: 変動遺伝子解析に関するファイルについての説明です。
- Help
本マニュアルを確認いただけます。