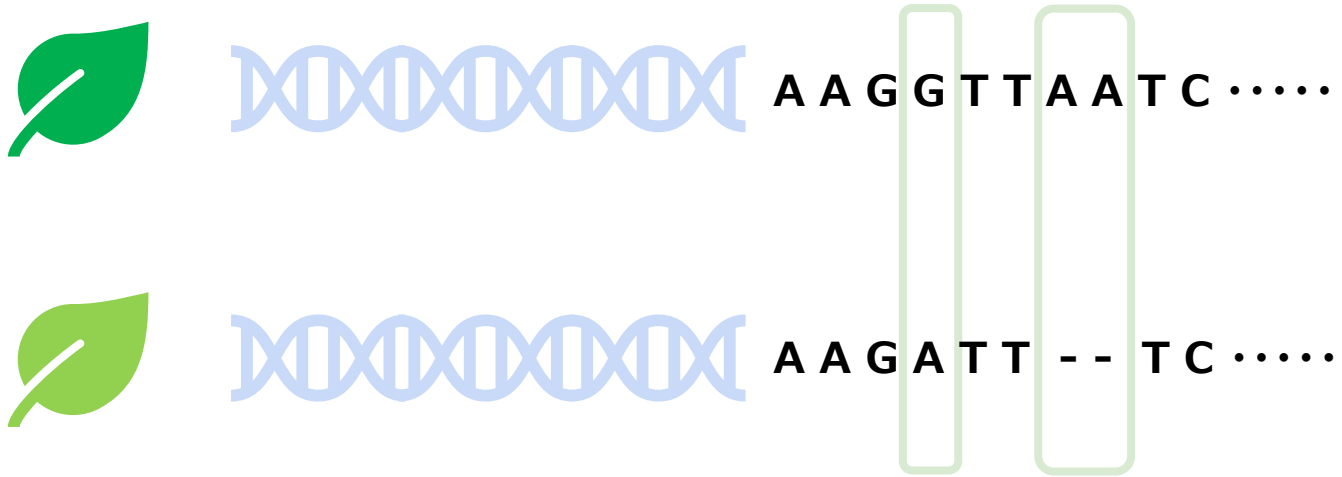
非モデル生物の遺伝子変異解析

フィルジェン株式会社
バイオインフォマティクス部(biosupport@filgen.jp)

2023.06

はじめに

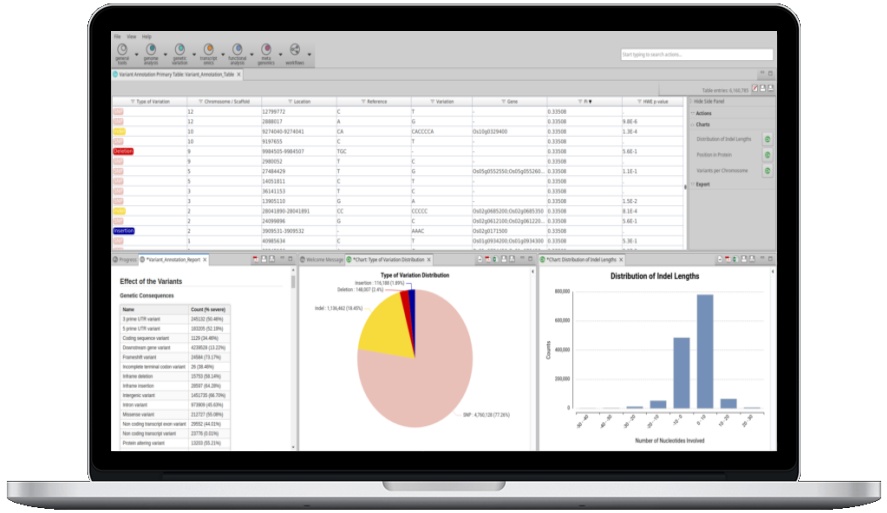
- 変異解析は配列データからバリアント（SNP、挿入、欠失など）を同定するプロセスである。
- 変異の原因や表現形への影響を理解することができるため、ゲノミクス研究において基本的な解析。



- モデル生物では十分に発展しているが、非モデル生物では、一貫した変異解析パイプラインのための十分なガイダンスが不足している。
- 非モデル生物にはモデル生物に利用可能な遺伝情報が欠けていることが多く、遺伝的変異解析の実行が困難。

OmicsBox

OmicsBoxの遺伝子変異解析機能



- 倍数性に対応したVariant Calling
- バリアントのフィルタリング
- バリアント アノテーション

その他

- ゲノムアライメント
- 遺伝子の機能予測

必要なファイル

BAM	FASTA	GTF
アラインメントされた配列ファイル	リファレンスゲノム	アノテーション ファイル

解析ワークフロー

QC・トリミング

- NGSより出力された生データが良好か、下流分析に影響する問題がないか確認。
  FastQCとTrimmomaticツールを統合

アライメント

- リード配列をリファレンスゲノムへマッピングする。
  BWA、Bowtie 2ツールを統合

Variant Calling

- アライメントファイルとリファレンスゲノムから、変異個所を検出する。
  BCFtools、Freebayesを統合

バリアントのフィルタリング

- 信頼性の高いバリアントを特定。

バリアントアノテーション

- バリアントに機能情報 (バリアントのコーディングや遺伝的影響など) を割り当てるプロセス
  Ensembl Variant Effect Predictor (VEP)を統合

QC・トリミング

・データが良好か、下流分析に影響する問題がないか確認

Welcome Message | FASTQ Quality Check (Dataset) | FASTQ Quality Check (ERR1948631_1.fastq) | FASTQ Quality Check (clean_ERR1948631_1.fq) | *Chart: Adapter Conten

FASTQ Quality Check

Name: Dataset

Overall Results

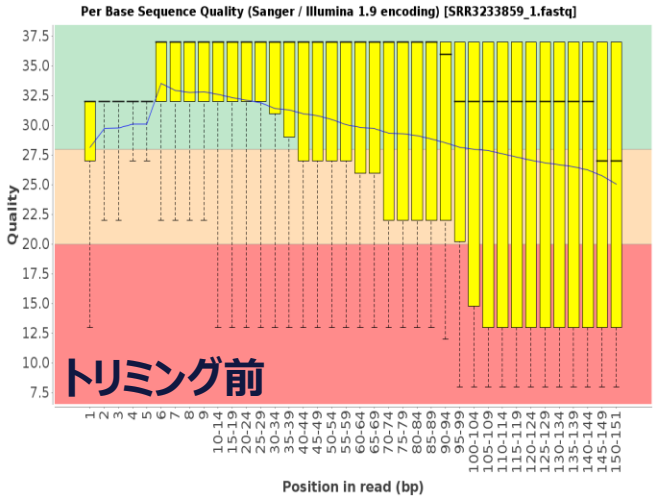
Name	Per Base Sequence Quality	Per Sequence Quality Scores	Per Base Sequence Content	Per Sequence GC Content	Per Base N Content
ERR1948631_1.fastq	PASS	PASS	FAIL	PASS	PASS
clean_ERR1948631_1.fq	PASS	PASS	FAIL	PASS	PASS

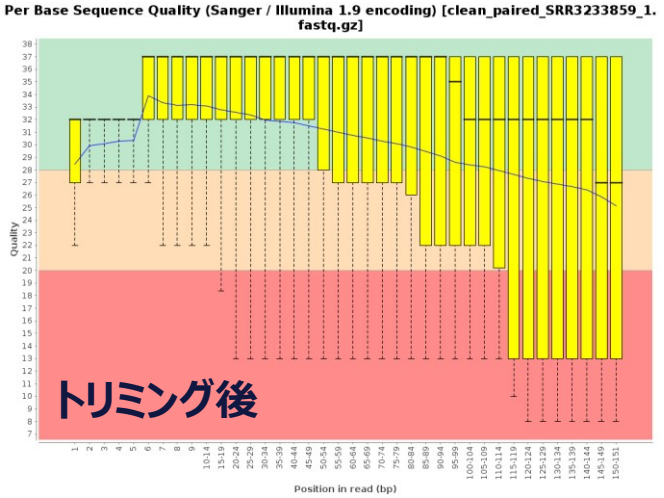
Name	Sequence Length Distribution	Adapter Content	Overrepresented Sequences	Sequence Duplication Levels	Report
ERR1948631_1.fastq	PASS	FAIL	WARNING	FAIL	
clean_ERR1948631_1.fq	WARNING	PASS	WARNING	FAIL	

The FASTQ quality check task is performed by nine analysis modules. The table above provides a quick evaluation of whether the results of each module seem entirely normal (pass), sightly abnormal (warning) or very unusual (fail). Note that these evaluations must be taken in the context of what is expected from the library. For example, some experiments may be expected to produce libraries which are biased in particular ways. Therefore, the summary evaluations should be treated as pointers that guide the preprocessing of the libraries.

✓解析が終了するとレポートが作成

正常（PASS）
わずかに異常（WARNING）
異常（FAIL）





シーケンスデータの品質をすばやく評価

レポートのアイコンをクリック→さらに詳細な結果を見ることが可能

アライメント

*Genome Analysis Module使用

・ショートを参照ゲノムに効率的にマッピングして、リードの起源となった正しい遺伝子座を特定する
→Variant Callingに必要なBAMファイルの作成が目的

Read Alignment (BWA) Results

Input 1: Reference Genome Sequences

assembly

Sequences	Minimum Length	Maximum Length	Average Length	Total Length
1	4,045,279	4,045,279	4,045,279	4,045,279

Input 2: FASTQ Files

File Name	Sample Name	Format	Sequencing
ERR2935851_1.fastq.gz, ERR2935851_2.fastq.gz	ERR2935851	FASTQ	Paired End

Results Overview

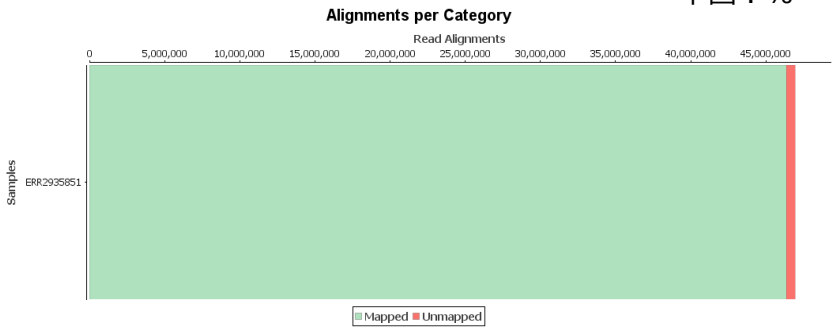
Globals

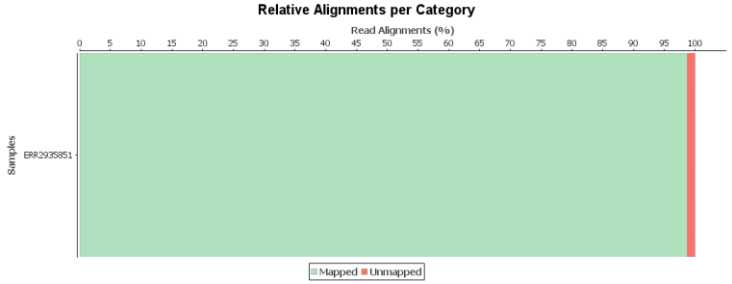
Sample	Total Alignments	Mapped	Supplementary	Unmapped	Duplicated Reads (estimated)	Duplication Rate
ERR2935851	46,952,430	46,358,042 / 98.734%	623,990 / 1.329%	594,388 / 1.266%	42,318,677 / 90.131%	99.4

どのくらいリードがマッピングされたか？

✓解析が終了するとレポートが作成

上図：リード長
下図：%



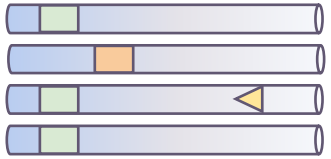


カテゴリーグラフ

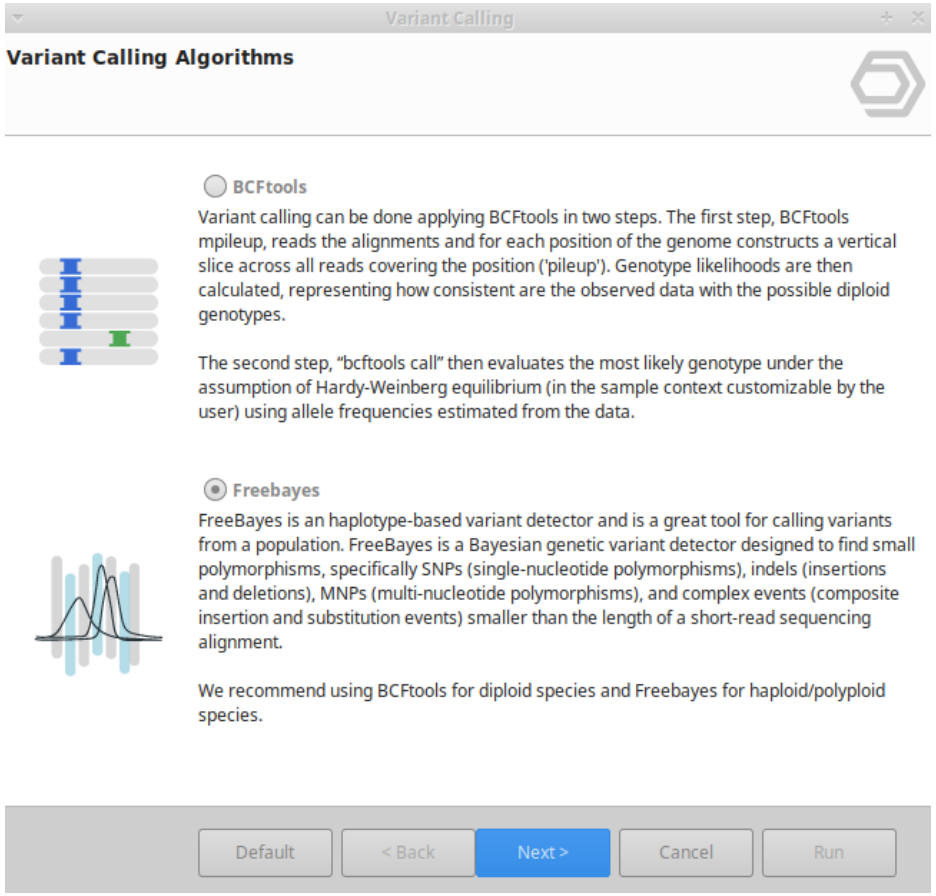
BAM

アラインメントされた配列ファイル

Variant Calling



・前項で作成したアライメントファイル（BAMファイル）とリファレンスゲノムファイルを使用して変異個所を検出する。



BCFtools

二倍体の生物向けのVariant Callingパッケージ

Freebayes

一倍体/倍数体の生物向けのVariant Callingパッケージ

データに適切なパッケージを選択可能

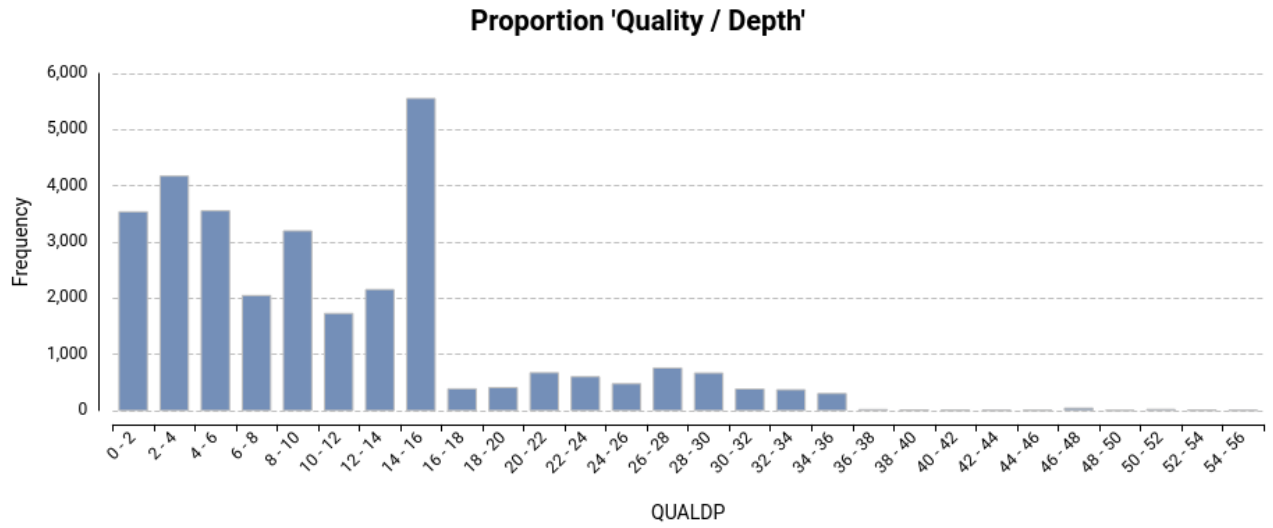
Variant Calling

Results

VCF File Saved as: /home/enrique/OmicsBoxWorkspace/freebayes_ew.vcf.gz

Type of variant	Frequency
SNP	3546
MIXED	2
MNP	182
INDEL	269

Number of alleles in a variant	Frequency
2	3989
3	9
4	1



✓解析が終了するとレポートが作成

✓さまざまな品質パラメーターの分布を示すいくつかのグラフ
この情報を参考に次項のフィルタリングを行う。

VCF

見つかったすべてのバリアントを含むVCF ファイル

バリアント フィルタリング

・信頼性の高いバリアントが特定され、誤って検出バリアントが削除する。

Variant Filtering Report

Input Data

`freebayes.vcf.gz`

Results

Number of variants before filtering	Number of variants after filtering	Percentage filtered
3960	1001	74.7222%

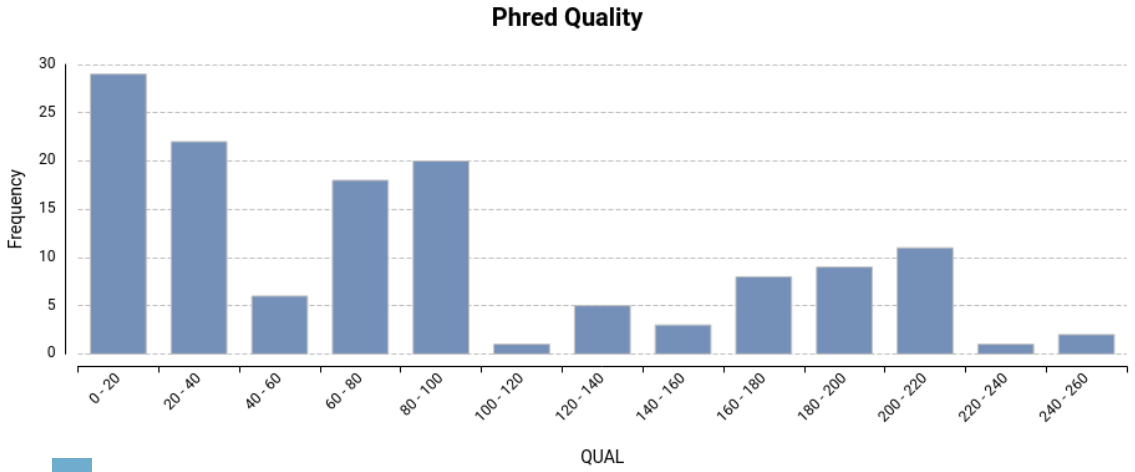
Parameters

Parameter	Value
Proportion 'Quality / Counts'	2.0
Raw Read Depth	5
Check Reads in Both Strands	true
Check if Reads are Balanced	true
Average Mapping Quality	50.0
Phred Quality	1.0
Remove Multiple Alleles	true

References

- Danecek P et al. (2021). Twelve years of SAMtools and BCFtools. *GigaScience, 10*(2).
- OmicsBox - Bioinformatics made easy. BioBam Bioinformatics (Version 3.0.23). March 3, 2019. www.biobam.com/omicsbox.

✓保持されているバリアントの数を表すレポートが作成



✓更新されたをグラフが作成される。この情報に基づいて、分析を続行するか、フィルタリング ステップのしきい値を調整するかを決定できる。

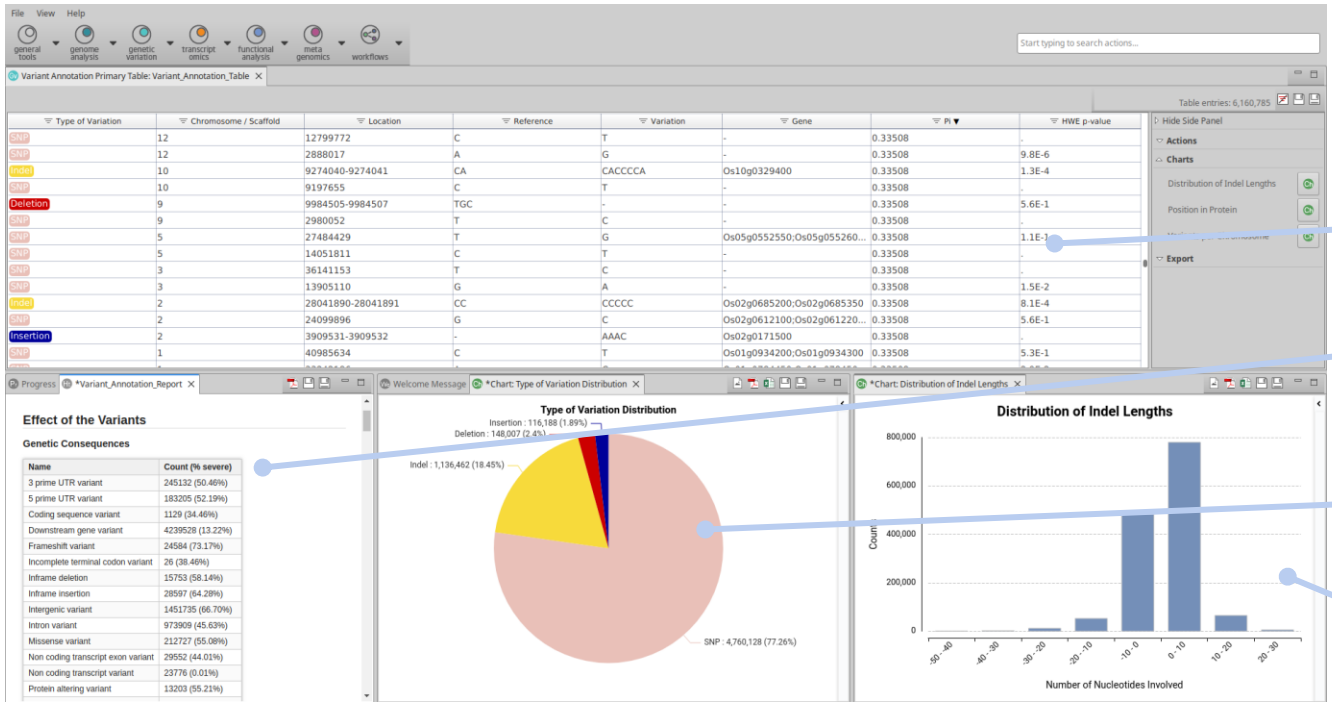
VCF

信頼性の高いバリアントを含むVCF ファイル

バリアントアノテーション

GTF

・バリアントに機能情報（バリアントのコーディングや遺伝的影響など）を割り当てるプロセスであり、ゲノム配列解析において重要なプロセス。
別途外部データベースなどからアノテーションファイル（GTFファイル）をダウンロードしておく必要がある。



✓次のような包括的な情報が作成

・ 各変異の位置、変異によって影響を受ける遺伝子、および2つの集団遺伝学値を示す表 。

・遺伝的およびコーディングの結果とサンプルのヘテロ接合性に関する情報を含むレポート。

・バリアントの種類の分布を示す円グラフ。

・品質管理チャート

バリアントアノテーション

バリアントによって影響を受ける遺伝子機能の名前 (または ID)

amino acid change
バリアントがタンパク質コード配列に影響を与える

Annotation Details

Gene: Os06g0130800

Feature ID	Feature Type	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Impact	Distance	Strand
Os06t0130800-01	Transcript	upstream_gene_variant	-	-	-	-	-	MODIFIER	2453	-1

Gene: Os06g0130900

Feature ID	Feature Type	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Impact	Distance	Strand
Os06t0130900-00	Transcript	frameshift_variant	66-71	65-70	22-24	EKA/EKGX	gAGAAAGcc/gAGAAAGGcc	HIGH	-	-1

Gene: Os06g0131001

Feature ID	Feature Type	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Impact	Distance	Strand
Os06t0131001-00	Transcript	frameshift_variant	567-572	567-572	189-191	GFL/GLSX	ggCTTTCTc/ggCCTTTCTc	HIGH	-	1

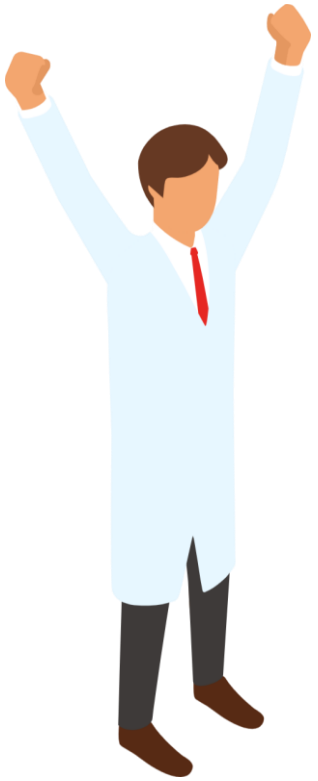
Gene: Os06g0131100

Feature ID	Feature Type	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Impact	Distance	Strand
Os06t0131100-01	Transcript	upstream_gene_variant	-	-	-	-	-	MODIFIER	2338	1
Os06t0131100-02	Transcript	upstream_gene_variant	-	-	-	-	-	MODIFIER	2360	1

✓バリアントの行を右クリックし、「*Show Annotation Details*」をクリックすると、そのバリアントの影響を受ける遺伝子ごとに1つのテーブルが表示されたレポートが開く。

OmicsBox のGenetic Variation

- アラインメントファイル、リファレンスゲノムゲノム、アノテーションファイルの3つのファイルのみで、あらゆる場面で簡単に遺伝子変異解析を行うことができる。
- 非モデル生物のバリアントのアノテーションに適している。
- Variant Callingとフィルタリングの実行は、コマンド不要のウィザードで簡単に操作できる。さらに、OmicsBoxは、信頼性の低いバリアントを除外し、最も信頼できるバリアントのみを選択してさらなる解析を行うことができる品質管理チャートとサマリーレポートを作成できる。



Information

弊社HP紹介ページ

お問い合わせ先：フィルジェン株式会社

TEL 052-624-4388 (9:00～17：00)

FAX　052-624-4389

E-mail: biosupport@filgen.jp