

鈴木 貴之、坊農 秀雅

広島大学 大学院統合生命科学研究科

## 要旨

- PubMedの文献を、遺伝子/生物種/実験手法（ゲノム編集ツール）のサブセットでまとめたデータベース
- 特定の遺伝子や生物種における、ゲノム編集研究の論文リストを得るときに有用です

### GEM: Genome Editing Meta-database

Extracted from PubMed Central, PubTator Central, NCBI gene

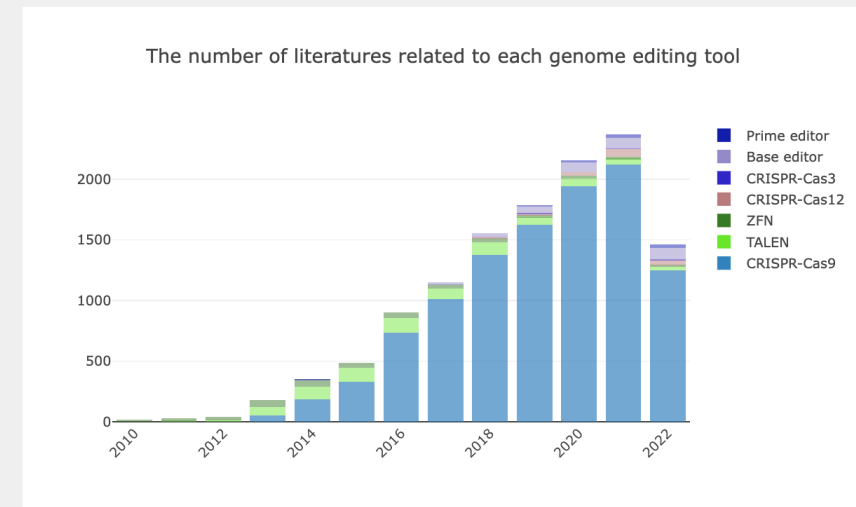

[About](#)
[News](#)
[Dataset](#)
[Figures](#)
[Bonohulab](#)

#### What is GEM?

GEM provides a collection of metadata extracted from PubMed

#### Browse by Publication Years

Please select the area on the figure below to see the desired dataset

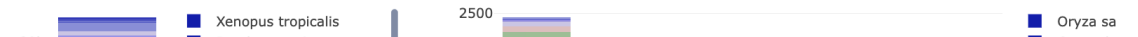


#### Browse by Species

Please select the area on the figure below to see the desired dataset

The number of genes studied by CRISPR-Cas9 for each specie

The number of genes studied by other tools for each specie



## Introduction

- ゲノム編集ツールが身近になったこともあり、今後は様々な生物種における様々な遺伝子の機能解析研究が進むことが予想されます。その際に、「ゲノム編集実験成功例がある生物種/遺伝子は何か」「逆に未研究の遺伝子は何か」というような情報に迅速にアクセスできると、新規研究の仮説を立てることに繋がります。
- 上記のような情報は文献の中に記載がありますが、アクセスが難しい状況にあります。遺伝子やゲノム編集ツールなどは表記揺れも多いため、網羅的に情報を収集することが困難です。
- よって、本研究では上記のような情報に容易/網羅的にアクセス/収集するためのデータベースを作成します。

まだ公開までは至っていないデータベースですが、  
現状の進捗を報告させていただきます。

1124

Search for keywords... Show 20 entries

tool	pubmed_id	publication title	pubdate	organism_name	genesymbol	editing type	gene_counts	biopro_id	RNA-seq	vector
CRISPR-Cas9	30905411	Deficiency of lysyl hydroxylase 2 in mice causes systemic endoplasmic reticulum stress leading to early embryonic lethality.	2019-03-21	Mus musculus	Plod2	Knock-Out	45	null	null	null
CRISPR-Cas9	30900179	Induced Neurons for the Study of Neurodegenerative and Neurodevelopmental Disorders.	2019-3-23	Homo sapiens	null	null	null	null	null	null
CRISPR-Cas9	30900175	One-Step Generation of Seamless Luciferase Gene Knockin Using CRISPR/Cas9 Genome Editing in Human Pluripotent Stem Cells.	2019-3-23	Homo sapiens	FMR1	Knock-in	825	null	null	null
CRISPR-Cas9	30900100	Multiple areas investigation reveals the genes related to vascular bundles in rice.	2019-03-21	Oryza sativa	null	null	null	null	null	null
CRISPR-Cas9	30898877	Leukodystrophy-associated	2019-03-21	Homo sapiens	POLR3A	Knock-Out	104	null	GSE118739	null
CRISPR-Cas9	30898877	Leukodystrophy-associated	2019-03-21	Homo sapiens	BCYRN1	Knock-Out	43	null	GSE118739	null
CRISPR-Cas9	30898838	Genome-Wide Interrogation of Human Cancers Identifies EGLN1 Dependency in Clear Cell Ovarian Cancers.	2019-03-21	Homo sapiens	EGLN1	Knock-Out	216	null	null	null
CRISPR-Cas9	30898720	Avoiding the off-target effects of CRISPR/Cas9 system is still a challenging accomplishment for genetic transformation.	2019-03-18	Homo sapiens	null	null	null	null	null	null
CRISPR-Cas9	30898719	CRISPR/Cas9-mediated genome editing of splicing mutation causing congenital hearing loss.	2019-03-18	Homo sapiens	SLC26A4	null	311	null	null	null
CRISPR-Cas9	30897091	A genetic switch for worker nutrition-mediated traits in honeybees.	2019-03-21	Apis mellifera	Fem	null	11	null	null	MLM3613
CRISPR-Cas9	30897091	A genetic switch for worker nutrition-mediated traits in honeybees.	2019-03-21	Apis mellifera	Dsx	null	6	null	null	MLM3613
CRISPR-Cas9	30895956	[A Plasmid-Expressed CRISPR/Cas9 System Suppresses Replication of HSV Type 1 in a Vero Cell Culture].	2018-08-28	Homo sapiens	null	null	null	null	null	null
CRISPR-Cas9	30894629	A reference collection of patient-derived cell line and xenograft models of proneural, classical and mesenchymal glioblastoma.	2019-03-20	Homo sapiens	MGMT	null	752	PB_JNA508446	null	lentCRISPR v2

- 約364の生物種におけるゲノム編集研究論文がある中の、**37の生物種**に対応しています。
- 学術論文内において記載のある、**ゲノム編集ツール名、生物種名、遺伝子名、Editing Type**（ノックアウトかノックインか）、利用された**ベクター**、が整理されています。
- Gene2pubmedを基に、それぞれの遺伝子が過去にどれだけ研究されているかを`**gene\_counts**`で示しています。
- 文献内（MethodsやData availabilityなど）に記載のある**BioProject**や**NCBI GEO**の**ID**を示しています。

## メタデータ取得方法

生物種名や遺伝子名

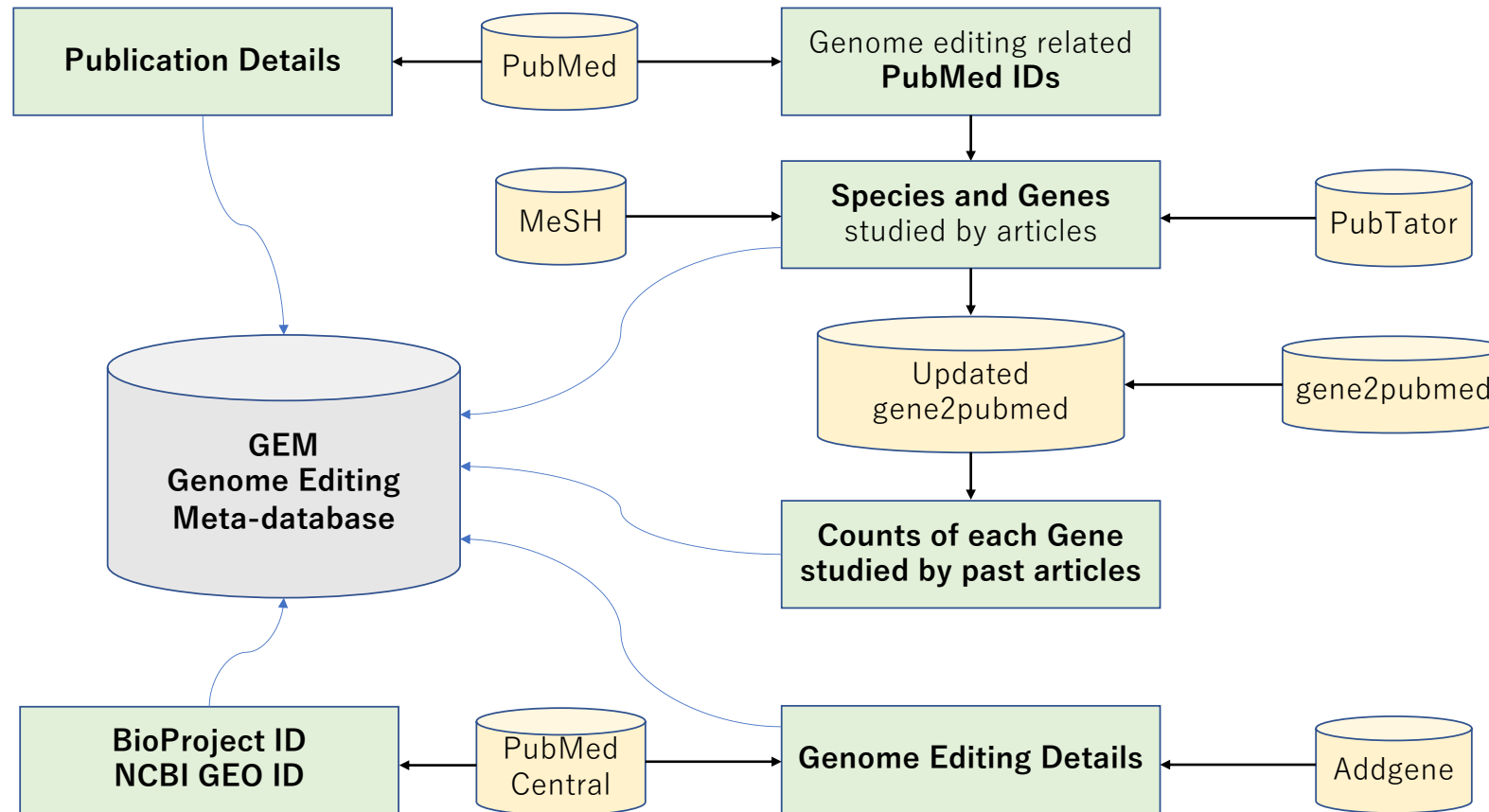
MeSHもしくはPubTator\*の情報を基に取得

ゲノム編集ツール名や  
Editing typeについて

文献のMethods/Results/Title/Abstractや  
PubMedのキーワードリストを基に、正規表現  
を利用して取得

\*PubTator: PubMedやPubMed Centralの文献に、遺伝子や生物種などの生物医学的概念を自動的にアノテーションするウェブベースのシステム。 <https://doi.org/10.1093/nar/gkz389>

## メタデータ 取得方法



## 結果

ゲノム編集関連の 文献数 (PMC)	探索した 生物種数	エントリー数
<b>33,011</b>	<b>37</b> /364	<b>39,996</b>

同じ文献内で、複数の  
- ゲノム編集ツール  
- 生物種  
- 遺伝子  
を研究している場合は、  
それぞれを別エントリーと  
して数えています。

## 類似データベース

- **Genome Targeting Catalogue (GTC) by EBI :**
  - CRISPR/Casシステムを利用した実験でターゲットとなった遺伝子のデータベース
  - キュレーションされたデータのみが登録されている
  - 2022年8月30日時点で、313エントリー (47生物種)が登録
- **European Sustainable Agriculture Through Genome Editing (eusage) :**
  - ゲノム編集を利用して作られた作物のデータベース
  - マニュアルキュレーションで作成
  - 2022年8月30日時点で583エントリーが登録



## Limitation

- PubMedもしくはPMCで公開されている論文のみに絞ってメタデータ取得を行なっている。
- PubTator Centralや正規表現を利用した情報抽出を行なっているため、データセットの内容に細かいエラーが存在する。

## Positive Control

- 取得したCRISPR-Cas9の論文は全て2013年以降（CRISPR-Cas9の最初の発表が2013年）、Prime editorの論文は全て2019年以降（Prime editorの最初の発表は2019年）
- 正しい情報が取得できていることを確認した論文の例
  - PMID: 27733139（豆の種子油の向上を目的にFAD3等の遺伝子をターゲットにしたゲノム編集実験）
  - PMID: 24179142（ゼブラフィッシュにおけるNHEJを利用したノックインゲノム編集実験）
  - PMID: 25434822（ゲノム編集を利用したDMD治療の研究）
  - PMID: 27050479（2016年のニワトリでのCRISPR-Cas9成功報告の論文）

## 今後の予定

- エラーのメタデータを最大限減らすためのスクリプト見直し
- データセットの全体像を確認するためのデータ可視化
- データセットを効果的に利用するためのインターフェース作成