DDBJ Annotated/Assembled Sequences

登録ファイル形式

配列ファイル

配列ファイルは、全登録データの配列を FASTA に類似した形式で記述したテキストファイルです。 配列ファイルは、1つの配列データは、”>” で始まる1行のヘッダ行と、2行目以降の実際のシーケンス文字列で構成されます。DDBJ では、エントリ間は配列情報終了フラグ (//) で区切ります。

例: 配列ファイル

```
>CLN01  <-- 1件目のエントリ名
ggacaggctgccgcaggagccaggccgggagcaggaagaggcttcgggggagccggagaa
ctgggccagatgcgcttcgtgggcgaagcctgaggaaaaagagagtgaggcaggagaatc
gcttgaaccccggaggcggaaccgcactccagcctgggcgacagagtgagactta
//      <-- 配列情報終了フラグ
>CLN02  <-- 2件目のエントリ名
ctcacacagatgcgcgcacaccagtggttgtaacagaagcctgaggtgcgctcgtggtca
gaagagggcatgcgcttcagtcgtgggcgaagcctgaggaaaaaatagtcattcatataa
atttgaacacacctgctgtggctgtaactctgagatgtgctaaataaaccctctt
//      <-- 配列情報終了フラグ
```

書式と構文

必ず、UME または Parser を用いて、配列ファイルとアノテーションファイルの書式をご確認ください。

- ベクター，リンカー，アダプターなどの配列（technical readと呼びます）は必ず除去してください。ただし、ベクターなどの配列自体を報告する場合は、除去する必要はありません。
- 特殊なケースを除き，最初(5’端)と最後(3’端)の塩基は n にならないように，未端の n は除去してください。また、特に EST などの場合，シークエンサ出力をそのまま送付するのではなく、末端の信頼できない出力を削除するなど精査してください。
- エントリ名は行頭の「>」に続けて、 space, “ double-quote, = equal, | pipe, > greater-than, [] angled brackets, ¥ back-slash を含まない半角英数字 32 文字以内で記載してください。
- エントリ名はエントリ毎にユニークな文字列を記載してください。clone 名，isolate 名といった個々のエントリによって異なる名称の使用が一般的です。
- 配列ファイルとアノテーションファイルの各エントリは、同一のエントリ名により対応づけます。アノテーション情報と配列は、対応するエントリに同じエントリ名をつけ、同じ順番になるようにそれぞれ入力してください。入力されたエントリ順にアクセッション番号を発行いたします。
- 塩基配列には a，t，g，c 以外にも、必要に応じて、各種核酸コードが使用可能です。
- 終端子として配列情報終了フラグ(//)を必ず入力してください。
- 途中にスペース、空行が入らないようにしてください。
- CON に該当する場合は、配列ファイルは AGP ファイルで代替することができます。

アノテーションファイル

アノテーションファイルは、全登録データの登録者, REFERENCE、Feature/Qualifierの情報等を記述した、Entry, Feature, Location, Qualifier, Value の５列からなるタブ区切りテキストファイルです。
スクリプト、(MS Excel などの) 表計算ソフト、テキストエディタ等で作成が可能です。

例:アノテーションファイル （入力必須項目）

Entry	Feature	Location	Qualifier	Value
COMMON	SUBMITTER		ab_name	Robertson,G.R.
			ab_name	Mishima,H.
			contact	Hanako Mishima
			email	mishima@ddbj.nig.ac.jp
			phone	81-55-981-6853
			fax	81-55-981-6853
			phext	3207
			institute	National Institute of Genetics
			department	DNA Data Bank of Japan
			country	Japan
			state	Shizuoka
			city	Mishima
			street	Yata 1111
			zip	411-8540
	REFERENCE		title	Mouse Genome Sequencing
			ab_name	Robertson,G.R.
			ab_name	Mishima,H
			year	2017
			status	Unpublished
	COMMENT		line	Please visit our website
			line	URL: http://www.ddbj.nig.ac.jp/
CLN01	source	1..12297	organism	Mus musculus
			mol_type	genomic DNA
			clone	PC0110
			chromosome	8
	CDS	join(<1..456,609..879,1070..1213)	product	protein kinase
			codon_start	2
CLN02	source	1..12393	organism	Mus musculus
			mol_type	genomic DNA
			clone	PC0210
			chromosome	8
	CDS	9365..9640	product	hypothetical protein

書式と構文

必ず、UME または Parser を用いて、配列ファイルとアノテーションファイルの書式をご確認ください。

Entry

登録ファイル形式：配列ファイルで示した配列ファイルと対応するエントリ名を入力してください。

エントリ名を入力後、次のエントリが始まる行までは、Entry カラムには何も入力しないでください。

Feature

Biological feature と 独自に規定された DDBJ 登録用 feature の 2 つのタイプがあります。各 Feature の記載方法については以下で解説します。

Feature 入力後、次の Feature が始まる行までは、Feature カラムには何も入力しないでください。

Location

Biological feature と PRIMARY_CONTIG で Feature の記載に隣接するカラムにのみ、記載が必要です。

Qualifier

各行に 1つ記載します。記載可能な Qualifier は Feature に依存します。詳細は以下で解説します。

Value

Qualifier に依存します。各 Qualifier の説明に従って記載してください。

その他

アノテーションファイルでは、空行が存在した時点でファイルの終わりと判断されます。従って、複数エントリを入力する場合は、登録する全てのエントリの入力が終わるまで、途中に空行を作らずに入力してください。

Biological Feature 記載に関する参照先

名称	更新日	備考
Feature Table Definition	2016/11/17	version 10.6
Feature/Qualifier 対応一覧表	2016/11/09	
登録の見本	2014/11/27	DDBJ フラットファイル中の feature の記載例

共通情報 COMMON

COMMON の入力について

- アノテーションファイルでは全てのエントリに共通な情報を入力するために COMMON というエントリ名を使用することができます。
- COMMON エントリに記載された情報はデータベースに読み込まれる際に全てのエントリに反映されます。
- 通常 COMMON は SUBMITTER/REFERENCE/COMMENT 等で使用しますが, Feature 以下(Feature, Location, Qualifier, Value)の情報が全てのエントリに共通であれば、Biological feature でも記載できます。

COMMON の活用

location に使用可能なメタ塩基番号'E'

例: COMMON に rRNA feature を記載

Entry	Feature	Location	Qualifier	Value
COMMON	rRNA	<1..>**E**	product	16S rRNA

配列長が異なるために Location が異なることを除けば、Feature 以下の Qualifier, Value の情報が全てのエントリで共通に記載可能なケース（例：rRNA 部分配列による系統解析など）があります。

そのような場合には、COMMON エントリに Feature を記載し location には、最後の塩基番号の代わりにメタ塩基番号として、**E** を記載することにより、全てのエントリに共通となる Feature を COMMON エントリに記載することが可能です。

clone, submitter_seqid, note, ff_definition に使用可能なメタ表記 '@@[entry]@@'

例: COMMON に source feature を記載

Entry	Feature	Location	Qualifier	Value
COMMON	source	1..**E**	organism	Homo sapiens
			mol_type	genomic DNA
			submitter_seqid	**@@[entry]@@**
			ff_definition	**@@[organism]@@** DNA, **@@[submitter_seqid]@@**

Location および clone 名や contig 名を除けば、Feature: source の Qualifier, Value の情報が全てのエントリで共通に記載可能なケース（例： EST, GSS, TSA, TLS, WGS, WGS scaffold (CON division)など）があります。

そのような場合、エントリ名に clone 名 または contig 名を使用する場合に限り、Feature: source を COMMON エントリに記載することが可能です。

- Location には、最後の塩基番号の代わりにメタ塩基番号として、**E** を記載します。
- 例に示した **@@[entry]@@** の形式で記載すると、配列ファイルから引用したエントリ名に置換されます。 **@@[entry]@@** を記載は clone, submitter_seqid, note, ff_definition の Value に限定しています。

SUBMITTER

例: アノテーションファイル内の SUBMITTER　（入力必須項目）

Entry	Feature	Location	Qualifier	Value
COMMON	SUBMITTER		ab_name	Robertson,G.R.
			ab_name	Mishima,H.
			consrtm	Mouse Genome Consortium
			contact	Hanako Mishima
			email	mishima@ddbj.nig.ac.jp
			url	http://www.ddbj.nig.ac.jp
			phone	81-55-981-6853
			fax	81-55-981-6853
			phext	3207
			institute	National Institute of Genetics
			department	DNA Data Bank of Japan
			country	Japan
			state	Shizuoka
			city	Mishima
			street	Yata 1111
			zip	411-8540

SUBMITTER で使用する Qualifier のリスト

Qualifier	Value 使用可能文字(注意事項)	Value 文字上限数
ab_name (登録者名)	英, .[period], ,[comma], -[hyphen], ‘ [single quote as apostrophe]	64
contact (コンタクトパーソン)	英, .[period], ,[comma], -[hyphen], ‘ [single quote as apostrophe], [space] (first, middle, last name の順で間に space を入れて入力)	first(64),middle(128), last(64)
consrtm (コンソーシアム名)	英, 数, [space], -[hyphen], ‘ [single quote as apostrophe], . [period], _[underscore], .[comma], ( ) # & @ / ; : + *	255
email	英, 数, @, .[period], -[hyphen], _[underscore]	64
	[space] 以外	255
phone, fax, phext	数, -[hyphen] (国番号の頭に + はつけない)	16
institute, department	[back-slash], ` [back-quote] 以外	255
country, state	英, 数, [space], -[hyphen], ‘[single quote as apostrophe], . [period], _[underscore], ,[comma], ( ) # & @ / ; : + *	32
city	英, 数, [space], -[hyphen], ‘[single quote as apostrophe], . [period], _[underscore], ,[comma], ( ) # & @ / ; : + *	64
street	英, 数, [space], -[hyphen], ‘[single quote as apostrophe], . [period], _[underscore], ,[comma], ( ) # & @ / ; : + *	255
zip	英, 数, -[hyphen]	16

SUBMITTER の書式

- SUBMITTER は各エントリに一件必ず入力していただく必要がありますが、全件共通の SUBMITTER を入力する場合には COMMON エントリに入力してください。
  エントリ毎に異なる SUBMITTER を入力したい場合には個々のエントリに記載してください。なお、COMMON エントリに SUBMITTER を記載した場合は、他の全てのエントリで SUBMITTER を使用することはできません。
- SUBMITTER の Qualifier: ab_name には複数の登録者を入力できます。DDBJ のフラットファイルには、ここで入力された順番に登録者が記載されます。複数の登録者の記載を強く推奨しています。
  登録者の中から一名、コンタクトパーソン を Qualifier: contact で再度指定してください。contact の Value には、full name を記載してください。
- Qualifier: ab_name の Value には、論文等の著者名に準ずる形式で、氏名の略記を記載してください。

  形式:

  last name[comma]first name の頭文字[period]middle name の頭文字[period]

例:

Miyashita,Y.

Robertson,G.R.

形式によって (氏名にハイフンを含む等)、Parser によるチェックで WAR レベルのメッセージが表示されることがありますが、記述内容に問題がない場合は、そのまま記載可能です。

- ab_name 以外の Qualifier の Value には、コンタクトパーソンの情報をそれぞれ一件ずつしか入力できません。複数の研究機関の情報を入力したい場合には、別途、ご連絡ください。

REFERENCE

例: アノテーションファイル内の REFERENCE （入力必須項目）

Entry	Feature	Location	Qualifier	Value
	REFERENCE		title	Sequence and analysis of mouse ch.8
			ab_name	Robertson,G.R.
			ab_name	Mishima,H.
			status	Published
			year	2003
			journal	Nature
			volume	8
			start_page	15
			end_page	20

REFERENCE で使用する Qualifier のリスト

Qualifier	Value 使用可能文字(注意事項)	Value 文字上限数
title (論文のタイトル)	[back-slash], ` [back-quote] 以外	255
ab_name (著者名)	英, .[period], ,[comma], -[hyphen], ' [single quote as apostrophe]	64
consrtm (コンソーシアム名)	英, 数, [space], -[hyphen], ' [single quote as apostrophe], .[period], _[underscore], ,[comma], ( ) # & @ / ; : + *	255
status	以下の何れか Unpublished, In press, Published	-
year	数(西暦4桁)	4
journal	[back-slash], ` [back-quote] 以外 (PubMed type abbreviation を入力)	128
volume, start_page, end_page	英, 数, -[hyphen]	8

REFERENCE の書式

- REFERENCE は各エントリに最低１つ、必須となります。
- **Qualifier: ab_name の Value には、論文等の著者名に準ずる形式で、氏名の略記を記載してください。**

  形式:

  last name[comma]first name の頭文字[period]middle name の頭文字[period]

  例:

  Miyashita,Y.

  Robertson,G.R.

  形式によって (氏名にハイフンを含む等)、Parser によるチェックで WAR レベルのメッセージが表示されることがありますが、記述内容に問題がない場合は、そのまま記載可能です。
- status が “In Press” の場合は、Qualifier: journal も必須となります。
- status が “Published” の場合は、Qualifier: journal, volume, start_page, end_page も必須となります。
- 今後論文に投稿する予定のない場合にも、status を “Unpublished” として REFERENCE を入力してください。

- journal には雑誌の PubMed type abbreviation を入力して下さい。
- REFERENCE を複数入力する場合には、登録する塩基配列を掲載する予定の論文、あるいは、既に掲載されている論文情報を REFERENCE の筆頭に入力し、参考文献はそれ以降に入力してください。
- 登録する塩基配列に関する REFERENCE 情報が全件に共通する場合は、COMMON エントリに入力してください。エントリ毎に異なる参考文献の情報を入力したい場合には個々のエントリに記載してください。
- COMMON エントリと個々のエントリの双方に REFERENCE を入力した場合には、COMMON の情報から順に、フラットファイル上に反映されます。

DATE

例: アノテーションファイル内の DATE と hold_date

Entry	Feature	Location	Qualifier	Value
COMMON	DATE		hold_date	20231125

DATE の書式

- DATE、hold_date は必ず COMMON エントリに入力してください。公開予定日が異なる場合には、ファイルを公開予定日毎に分けて作成してください。
- DATE にはデータの公開予定日(hold_date)を年月日の順で、半角数字 8 桁(例：20231125)で入力してください。
- - [hyphen] や / [slash] などの区切り文字を使用した場合はエラーとなります。
- 登録作業後、データの即時公開をご希望の場合には、DATE を入力しないでください。
- 公開予定日をご指定いただいた場合は、データ公開原則に基づいて、公開作業を行ないます。

COMMENT/ST_COMMENT

例: アノテーションファイル内の COMMENT と ST_COMMENT

Entry	Feature	Location	Qualifier	Value
	COMMENT		line	This clone was obtained at our laboratory.
	COMMENT		line	Please visit our web site.
			line	URL:http://www.ddbj.nig.ac.jp
	ST_COMMENT		tagset_id	Genome-Assembly-Data
			Assembly Method	GS De Novo Assembler v. 2.0
			Assembly Name	Mmus_1.0
			Genome Coverage	50x
			Sequencing Technology	454 GS FLX; ABI 3730

※ COMMENT には “一般 COMMENT” と “structured COMMENT” があります。詳細は以下をご覧ください。

COMMENT (一般 COMMENT) の書式

- 一般 COMMENT は必要に応じて登録者が自由な記述形式で内容を入力することができます。
- COMMENT は DDBJ フラットファイル上では 60 文字(スペースを含む)で自動的に改行されますが、任意の位置で改行したい場合には、Qualifier: line を指定して改行位置で Value を分けてください。
- Qualifier: line の Value には、 [back-slash] 以外の文字を使用可能です。
- 全件共通の COMMENT は COMMON エントリに入力してください。エントリ毎に異なる COMMENT を入力する場合には個々のエントリに記載してください。
- 内容の異なる COMMENT を入力する場合には、COMMENT と COMMENT の間に空行を入れるため、COMMENT Feature をそれぞれに作成してください。
- COMMON エントリと個々のエントリの双方に COMMENT を入力した場合には、COMMON から順に、フラットファイル上に反映されます。また、複数の COMMENT を入力した場合は、アノテーションファイルに入力した順番でフラットファイル上に反映されます。
- EST の場合、特殊な COMMENT の記載が必要なことがあります。

ST_COMMENT (structured COMMENT) の書式

- ST_COMMENT は一定のルールに従って構造化された COMMENT (structured COMMENT) を記載するための feature です。
- ST_COMMENT はユーザー定義も可能ですが、Genome Project (WGS も含む)、Transcriptome Project (TSA も含む) などの登録には既定書式があり、記載する必要があります。
- ST_COMMENT はデータセット名 (tagset_id) と項目名 (ユーザー定義 Qualifier) 、各項目の値 (Value)で構成されます。
- Structured COMMENT の開始行ではQualifier に tagset_id 、Value に COMMENT のタイトルを入力します。
  Genome Project の場合は tagset_id に Genome-Assembly-Data を入力します。
  Transcriptome Project の場合は tagset_id に Assembly-Data を入力します。
- 項目名を Qualifier として入力します。各項目に対応する具体的な内容を Value に入力します。
- Genome-Assembly-Data で使用する Qualifier のリスト （入力必須項目）

Qualifier	説明	備考
Assembly Method	アセンブルに使用したソフトウェア名とそのバージョン。必須。	必ずソフトウェアのバージョン番号を“ v. “直後に記載して下さい(例 Velvet v. 2.0)。
Assembly Name	ゲノムアセンブリの名称・バージョン。 真核生物の場合は必須。	推奨書式： [organism の種名 (or 一般名)] + [version 数値] （例 Btau_4.0）
Genome Coverage	ゲノム配列決定の深度、被覆度換算。必須(例 125x)。	Coverage不明時には “Unknown” を記載して下さい。
Sequencing Technology	配列解析に使用したシークエンサー。必須。	複数のシークエンサーが使われたときはセミコロンと半角スペースで挟んで記載してください(例 454 GS FLX; ABI 3730)。

- Assembly-Data で使用する Qualifier のリスト （入力必須項目）

Qualifier	説明	備考
Assembly Method	アセンブルに使用したソフトウェア名とそのバージョン。必須。	必ずソフトウェアのバージョン番号を“ v. “直後に記載して下さい(例 Velvet v. 2.0)。
Assembly Name	アセンブリの名称・バージョン。	推奨書式： [organism の種名 (or 一般名)] + [version 数値] （例 Btau_4.0）
Coverage	配列決定の深度、被覆度換算(例 125x)。	Coverage不明時に “Unknown” を記載可能です。
Sequencing Technology	配列解析に使用したシークエンサー。必須。	複数のシークエンサーが使われたときはセミコロンと半角スペースで挟んで記載してください(例 454 GS FLX; ABI 3730)。

- 記載の可否や内容等については登録毎に個別に対応しますので、MSS の担当者にお問い合わせください。

Biological Feature

例: アノテーションファイル内の source と CDS feature（入力必須項目）

Entry	Feature	Location	Qualifier	Value
	source	1..12297	organism	Mus musculus
			mol_type	genomic_DNA
			chromosome	8
			clone	PC0110
	CDS	join(<1..456,609..879,1070..1213)	product	protein kinase
			codon_start	2
	rRNA	1279..3000	product	18S rRNA
	CDS	complement(join(3213..4981,9901..11677))	gene	tbpA
			product	TATA-box binding protein

※ Biological feature の定義、記述方法の詳細については、Feature Table Definitionをご参照ください。

Feature/Location/Qualifier の書式

- Feature Table Definition では、各 Qualifier の前に / [slash] が記述されておりますが、アノテーションファイルでは / を入力しないでください。
- source と organism、mol_type は各エントリに最低１つ、必須となります。

- Location の記載ルールは、Location の記述法を ご参照ください。
- 各 Feature で使用可能な Qualifier は Feature/Qualifier 対応表 にて確認できます。 一部の Feature には、入力必須 Qualifier が指定されています。 対象の Feature で、Mandatory qualifier と指定されているものは必ず入力してください。 大文字と小文字の区別, _ [underscore] の使用も対応表の表記に従ってください。
- あわせて、アノテーションファイルのサンプル と 登録の見本も ご参照ください。
- CDS の記載に際しましては、タンパク質コード配列; CDS feature についてを ご参照ください。
- CDS feature を含むデータは、必ず、UME または transChecker を用いてアミノ酸翻訳をご確認ください。

Value の書式

- 使用可能な文字種は Qualifier に依存します。詳細は Feature Table Definitionおよび、Feature/Qualifier の対応一覧表をご参照ください。
- Value type に従い、各 Qualifier で指定されている文字種を使用して、正しく入力してください。

DIVISION

DIVISION は、登録データが CON / ENV / EST / GSS / HTC / HTG / STS / SYN / TSA のいずれかに該当することを示します。

例: アノテーションファイル内の DIVISION

Entry	Feature	Location	Qualifier	Value
COMMON	DIVISION		division	EST

DIVISION の書式

- Qualifier : division の Value にdivision の名称を示すアルファベット３文字を大文字で入力してください。
- DIVISION は、原則として COMMON エントリに入力してください。

DATATYPE

DATATYPE は、登録データが WGS, TLS, TPA, TPA-WGS の何れかに該当することを示します。

例: アノテーションファイル内の DATATYPE

Entry	Feature	Location	Qualifier	Value
COMMON	DATATYPE		type	WGS

DATATYPE の書式

- Qualifier: type の Value に WGS, TLS, TPA, TPA-WGS の何れかを入力してください。
- DATATYPE は COMMON エントリに入力してください。

KEYWORD

KEYWORD には、DIVISION と DATATYPE で示されたデータ種別を基本に、細分化した情報, 実験手法に関する情報などを、原則として、規定値で記載します。
INSDC が合意した KEYWORD 名と規定値、並びに各 KEYWORD 名の定義につきましては、INSDC agreed methodological keywords をご参照ください。

例: アノテーションファイル内の KEYWORD

Entry	Feature	Location	Qualifier	Value
	KEYWORD		keyword	ENV

データ種別ごとの keyword の Value入力必須項目

データ種別	keyword の Value	注意事項
WGS	WGS	WGS, scaffold CON の場合 もご参照ください。

データ種別	keyword の Value	注意事項
ENV	ENV	
EST	EST	
	その他	EST の場合 参照
HTC	HTC, その他	その他については、登録毎にご連絡いたします。
HTG	HTG, その他	phase に依存、登録毎にご連絡いたします。
GSS	GSS	
STS	STS	
TPA	TPA, Third Party Data	
	TPA:inferential or TPA:experimental	どちらか一方が必須
TSA	TSA, Transcriptome Shotgun Assembly	
TLS	TLS, Targeted Locus Study	
その他		登録毎にご連絡いたします。

KEYWORD の書式

- Qualifier: keyword の Value に該当する規定値を入力してください。
- 詳細な記載方法に関しましては、登録毎にご連絡いたします。

WGS, scaffold CON の場合

- WGS や WGS エントリを primary エントリに引用した scaffold 配列（CON エントリ）では、登録される塩基配列の完成度を示すため、次のいずれかを KEYWORD に記載してください。
  - STANDARD_DRAFT
  - HIGH_QUALITY_DRAFT
  - IMPROVED_HIGH_QUALITY_DRAFT
  - NON_CONTIGUOUS_FINISHED

  例: WGS draft genome（入力必須項目）

Entry	Feature	Location	Qualifier	Value
	KEYWORD		keyword	WGS
			keyword	STANDARD_DRAFT

EST の場合

- EST では、EST に加えて、以下のいずれかを keyword に必ず記載してください。
  - 5′ EST の場合 — 5′-end sequence (5′-EST)
  - 3′ EST の場合 — 3′-end sequence (3′-EST)
  - 上記を特定できない場合 — unspecified EST

  例: 5′ EST（入力必須項目）

Entry	Feature	Location	Qualifier	Value
	KEYWORD		keyword	EST
			keyword	5′-end sequence (5′-EST)

- 3′ EST では、登録される塩基配列が anti-sense 鎖側、sense 鎖側のどちらであるかを示すため、次のいずれかを COMMENT に記載してください。
  例: 3′ EST、anti-sense 鎖（入力必須項目）

Entry	Feature	Location	Qualifier	Value
	COMMENT		line	3′-EST sequences are presented as anti-sense strand.

  例: 3′ EST、sense 鎖（入力必須項目）

Entry	Feature	Location	Qualifier	Value
	COMMENT		line	3′-EST sequences are presented as sense strand.

HTG の場合

- HTG では、その配列決定の段階を示す keyword の記載を推奨しています。
  例I: 向きが不明な piece を含む場合（入力必須項目）

Entry	Feature	Location	Qualifier	Value
	KEYWORD		keyword	HTG
			keywrod	HTGS_PHASE1

Entry	Feature	Location	Qualifier	Value
			keyword	HTGS_DRAFT

例 II: 向きが不明な piece が含まない場合（入力必須項目）

Entry	Feature	Location	Qualifier	Value
	KEYWORD		keyword	HTG
			keyword	HTGS_DRAFT

DBLINK

DBLINK は、 BioProject ID、BioSample ID、Sequence Read Archive (DRA/ERA/SRA) 他, 特定データベースへのリンクを記載します。

例: アノテーションファイル内の DBLINK（入力必須項目）

Entry	Feature	Location	Qualifier	Value
	DBLINK		project	PRJDB12345
			biosample	SAMD90000000
			sequence read archive	DRR999000
			sequence read archive	DRR999001

DBLINK の書式

- 登録データが BioProject Database、BioSample Database に登録されている場合は、Qualifier: project の Value に BioProject ID、Qualifier: biosample の Value に BioSample ID を記載してください。
- 登録データが次世代シーケンサ由来のアセンブルで、raw reads が Sequence Read Archive に登録されている場合、Qualifier: sequence read archive の Value に対応する Run データのアクセッション番号を入力してください。
- BioProject Database, BioSample Database, Sequence Read Archive もご参照ください。

locus_tag

アノテーションが付加された全ゲノム規模の登録に関しましては、タンパク質産物 (CDS)、あるいは、転写産物(rRNA, tRNA など)を示す Biological feature に locus_tag を付加することを推奨しています。locus_tag prefix は事前に BioSample Database で BioSample ID を申請する際に取得して下さい。

source: ff_definition

ff_definition は、The DDBJ/EMBL/GenBank Feature Table: Definition には定義されていない DDBJ 登録専用 Qualifier です。必要な場合にのみ、 1 エントリに 1 つ記載します。

例: アノテーションファイル内の ff_definition

Entry	Feature	Location	Qualifier	Value
	source	1..516	organism	Mus musculus
			mol_type	mRNA
			ff_definition	**@@[organism]@@** mRNA, clone: **@@[clone]@@**
			clone	PC0110

ff_definition 記述フォーマット

データ種別	ff_definition記述フォーマット
WGS	@@[organism]@@ @@[strain]@@ DNA, @@[submitter_seqid]@@, [other information]
BAC/YAC genomic clones in unfinished phase (HTG)	@@[organism]@@ DNA, chromosome @@[map]@@, [BAC/YAC] clone: @@[clone]@@, *** SEQUENCING IN PROGRESS ***
BAC/YAC genomic clones in finished phase	@@[organism]@@ DNA, chromosome @@[map]@@, [BAC/YAC] clone: @@[clone]@@
EST	@@[organism]@@ mRNA, clone: @@[clone]@@, [other information]
EST	@@[organism]@@ cDNA, clone: @@[clone]@@, [other information]
GSS	@@[organism]@@ DNA, clone: @@[clone]@@, [other information]

データ種別	ff_definition記述フォーマット
STS	@@[organism]@@ DNA, @@[map]@@, [marker name], sequence tagged site
その他	登録毎にご連絡いたします。

source: ff_definition の書式

- Biological feature である source に Qualifier: ff_definition を入力します。
- ff_definition の記載内容は、DDBJ フラットファイル において DEFINITION 行に反映されます。詳細はサンプルとフラットファイルとの対応 をご参照ください。
- ff_definition の Value には、通常、同じ source feature 内にある他の Qualifier から Value を引用することが多いため、引用のためのメタ表記を用意しております。例に示したように **@@[organism]@@, @@[clone]@@** の形式で、Value を引用する Qualifier の名称を **@@[** と **]@@** で括り記載しておきますと、DEFINITION 行に反映する際に対象 Value に置換されます。
- 上記表に示した記述フォーマットを基本としますが、ff_definition の詳細な記載方法に関しましては、登録毎にご連絡いたします。

assembly_gap: Sequencing Gap Region

HTG に代表される大規模ゲノム配列やESTアセンブルによるトランスクリプトーム (TSA) 配列の登録などにおいて、アセンブル途上、難読領域であるなどの理由により生じる sequencing gap を示すために配列ファイルでは、配列中に 'n' を記載します。このとき、アノテーションファイルでは、その sequencing gap 領域を下記の要領で、assembly_gap feature を用いて示す必要があります。

例: アノテーションファイル内の assembly_gap（入力必須項目）

Entry	Feature	Location	Qualifier	Value
	assembly_gap	101..200	estimated_length	unknown
			gap_type	within scaffold
			linkage_evidence	paired-ends

assembly_gap: Sequencing Gap Region の書式

- assembly_gap feature は Biological feature の 1 つですが、特殊な書式になります。
- assembly_gap では、location に join, order, complement を使用することはできません。

長さが不明の場合
長さが判明していないギャップ領域については、登録者によって指定された一律の長さ (1000 bp未満の reasonableな長さ) の 'n' で記述する規則となっております。
また、Qualifier: estimated_length で Value に unknown と記載します。
但し、CON divisionではない transcriptome エントリ (TSA division など) には、Value に unknown を記載することができません。

長さが予測されている場合
長さが判明しているギャップ領域については、配列の相当位置に推定される長さの 'n' で記述する規則となっております。 また、Qualifier: estimated_length で Value に known と記載します。

TOPOLOGY

TOPOLOGY は登録塩基配列全体の形状が環状で、最初の塩基と最後の塩基が実際には連続している場合に記載する必要があります。
例：環状ウイルスゲノムの全長など

例: アノテーションファイル内の TOPOLOGY

Entry	Feature	Location	Qualifier	Value
	TOPOLOGY		circular	

TOPOLOGY の書式

- DDBJ フラットファイルでは、topology は LOCUS 行に反映されます。詳細はアノテーションファイルのサンプルを参照してください。

TPA/TSA: PRIMARY_CONTIG プライマリーエントリ引用

PRIMARY_CONTIG, entry, および primary_bases は、プライマリーエントリからの配列引用情報を記載するために設けられた TPA/TSA データ登録専用の Feature, Qualifier です。

例: アノテーションファイル内の PRIMARY_CONTIG

Entry	Feature	Location	Qualifier	Value
	PRIMARY_CONTIG	1..438	entry	ZZ000010.1
			primary_bases	1..438
	PRIMARY_CONTIG	377..696	entry	ZZ000011.1
			primary_bases	1..320
			complement	
	PRIMARY_CONTIG	590..1191	entry	ZZ000022.0
			primary_bases	1..601

PRIMARY_CONTIG feature で使用可能な qualifier

Qualifier	Value 記述時の注意事項
entry	引用するエントリのアクセッション番号を(バージョン番号とともに)入力する
primary_bases	引用したプライマリーシークエンスの位置情報 各配列の領域を入力する 例) 1..500
complement	引用するエントリが相補鎖である場合に入力が必要

TPA/TSA: PRIMARY_CONTIG プライマリーエントリ引用 の書式

- DATATYPE/type で TPA、もしくは DIVISION/division で TSA を指定しておく必要があります。
- PRIMARY_CONTIG には、引用後の結果として配列(TPA/TSA)上の位置情報、および、引用したプライマリーシークエンスの(バージョンの付いた)アクセッション番号とその位置情報を必ず入力していただきます。
- Location に join, order, complement を使用することはできません。同じ entry を引用する場合も location 単位で PRIMARY_CONTIG を複数記述してください。
- プライマリーシークエンスが DDBJ/EMBL-Bank/GenBank に登録されている場合は、バージョンの付いたアクセッション番号を記載します。引用したアクセッション番号のデータが、TPA/TSA データ登録時点では、まだ公開されていない場合は、バージョン番号は、0 [zero]と入力してください。
- 詳細はサンプルとフラットファイルとの対応を参照してください。
  - TPA (Third Party Data) : サンプル
  - TSA (Transcriptome Shotgun Assembly) : サンプル
  - TSA; assembled from short reads : サンプル

サンプルアノテーション

一般データ	タンパク質コード領域	CDS
	リボソーマル RNA	16S_rRNA
	ITS 領域 (Internal Transcribed Spacer)	ITS
	マイクロサテライトマーカー	Microsatellite marker
	ミトコンドリア	mtDNA
	ENV (環境サンプル)	ENV
ゲノムデータ関連	complete genome sequence (Bacteria)	complete_genome_BCT
	Finished level genome sequence with biological feature (Eukaryote)	Finished_genome_eukaryote
	WGS (Whole Genome Shotgun) without annotation	WGS
	WGS (Whole Genome Shotgun) with annotation	WGS_annotation
	WGS; piece of scaffold CON	WGS_piece_CON
	CON entries for WGS scaffold	WGS_scaffold
	MAGs (Metagenome-Assembled Genomes, MAGs) for Complete genome	MAGs_CompleteGenome
	MAGs (Metagenome-Assembled Genomes, MAGs) for Draft genome	MAGs_WGS

	AGP file for CON entries	AGP
	GSS (Genome Survey Sequences)	GSS
	HTG (High Throughput Genomic Sequences)	HTG
大量転写物配列関連	TSA (Transcriptome Shotgun Assembly); assembled from EST	TSA
	TSA; assembled from short reads without annotation	TSA_SRA_assemble_NoANN
	TSA; assembled from short reads with annotation	TSA_SRA_assemble_Ann
	EST (Expressed Sequence Tags)	EST
TLS (Targeted Locus Study)	TLS (Targeted Locus Study)	TLS
TPA (Third Party Data)	TPA (Third Party Data)	TPA
	TPA assembly (Third Party Data)	TPA-assembly_WGS
	TPA assembly (Third Party Data)	TPA-assembly
アノテーション:フラットファイル	タンパク質コード領域	ann2-ff

AGP ファイル

AGP ファイルは CON エントリの登録に必要です。 AGP ファイルは CON エントリを構築する際のピースエントリの順序、種類、方向等が記載された、9列からなるタブ区切りテキストファイルです。 スクリプト、(MS Excel などの) 表計算ソフト、テキストエディタ等で作成が可能です。

AGP ファイルの書式は、UCSC, EBI および NCBI により開発されました。

例： AGP ファイル

#1	2	3	4	5	6	7	8	9
scaffold1	1	1345	1	W	BZZZ01123456.1	1	1345	+
scaffold1	1346	2845	2	N	1500	scaffold	yes	align_genus
scaffold1	2846	4301	3	W	BZZZ01123457.1	1	1456	+
scaffold1	4302	4401	4	U	100	scaffold	yes	align_genus
scaffold1	4402	5631	5	W	BZZZ01123458.1	1	1230	-
scaffold2	1	650	1	W	BZZZ01123486.1	1	1345	+
scaffold2	651	750	2	N	100	scaffold	yes	align_genus
scaffold2	751	2980	3	W	BZZZ01123488.1	1	1230	-

書式と構文

AGPファイルは、UME (Utilities for MSS Error check)でチェックすることが可能です。

- AGP ファイルは 9 カラムで構成されています。
- タブ区切りテキスト形式で作成してください。
- 途中にスペース、空行が入らないようにしてください。
- # で始まる行はコメント扱いとなります。データには反映されません。ファイルの先頭に記載してください。

各カラムにおける記述内容（カラム 1 - カラム 5）

カラム	内容	入力項目・注意事項	
1	object	CONエントリ名 (chromosome, scaffold, contig 等に対する固有の名称) アノテーションファイルのエントリ名と対応するエントリ名を入力する	
2	object_beg	CON エントリにおける component/gap の開始位置	
3	object_end	CON エントリにおける component/gap の終了位置	
4	part_number	CON エントリを構成する component/gap の順序	
5	component_type	component の種類を示す規定値: A, D, F, G, O, P, W, N, U のいずれか	
		A	Active Finishing; finishing に向けて更新され得る段階
		D	Draft HTG; HTG phase1, phase2, あるいは不明な draft 段階 つまり finished レベルに達していない HTG

カラム	内容	入力項目・注意事項	
		F	Finished HTG; phase3, finished レベルの HTG
		G	Whole Genome Finishing
		O	Other sequence; WGS, HTG に該当しないもの
		P	Pre Draft
		W	WGS contig; ピースエントリが WGS エントリである場合
		N	サイズが特定・予測されている gap
		U	サイズ不明の gap、100 塩基とすること

* component: より大きな配列を構築するために使用される配列 (ピースエントリ)

6 以降のカラムは、カラム 5 の value に依存して記述内容が異なります。

各カラムにおける記述内容（カラム 6 - カラム 9）：カラム 5 が “N”と”U”以外の場合

カラム	内容	入力項目・注意事項	
6	component_id	component のアクセッション番号とバージョン番号、あるいは component のエントリ名	
7	component_beg	component の開始位置	
8	component_end	component の終了位置	
9	orientation	component の相対的な配列の方向。規定値は下記：	
		+	プラス、順鎖
		-	マイナス、相補鎖
		?	不明
		0	ゼロ、不明 (deprecated)
		na	irrelevant
		ただし、"?"、"0"、"na" も順鎖と扱う。	

* component: より大きな配列を構築するために使用される配列 (ピースエントリ)

- 長さが判明していないギャップ領域については、一律 100 個の n で記述する規則となっています。component_type の value に”U”、gap_length の value に “100” と記載します。
- カラム 5 が “N”あるいは”U”の場合、連続性の情報は gap_type および linkage の組み合わせで与えられます。 以下の表を参考にしてください。
  例: アノテーションファイル内の COMMENT と ST_COMMENT

gap_type	linkage	解説・注意事項
scaffold 内の gap: gap 前後の配列が 1 つの scaffold に収まる場合、連鎖している		
scaffold	yes	scaffold を分けずに記載すること gap 前後の配列が連鎖する証拠があることを示す
repeat	yes	scaffold を分けずに記載すること gap に未解消の繰り返し単位が存在し、前後の配列が連鎖する証拠がある場合は 'yes' とする
scaffold を分ける gap: gap 前後の配列が分かれた scaffold に それぞれ位置し、連鎖するか否か不明		
contig	no	scaffold を分けて記載すること gap 前後の配列が連鎖する証拠がなく、連鎖するか否か不明
repeat	no	scaffold を分けて記載すること gap に未解消の繰り返し単位が存在し、前後の配列が連鎖する証拠がない場合は 'no' とする
centromere short_arm heterochromatin telomer	no	scaffold を分けて記載すること これら生物学的 gap は chromosome に沿った scaffold の間に配置すること
使用禁止となる gap type と linkage の組み合わせ		
contig	yes	この組み合わせは使用禁止 もし、gap 前後の配列が連鎖する証拠があるならば、gap type は contig ではなく scaffold とすべきである
scaffold	no	この組み合わせは使用禁止 もし、gap 前後の配列が連鎖する証拠がないならば、gap type は scaffold ではなく contig とすべきである
centromere short_arm heterochromatin telomere	yes	この組み合わせは使用禁止 これら生物学的 gap は scaffold 内では使用しないこと

Related pages

MSS データファイル用チェックツール
Parser ユーザーマニュアル
validator エラーメッセージ
UME ユーザーマニュアル
transChecker ユーザーマニュアル
MSS 利用申し込み

検索

DDBJ Search
getentry
ARSA
TXSearch

解析

Vector Screening System
WABI (Web API for Biology)
DDBJ FTP Site

データベース

Annotated/Assembled Sequences (DDBJ)
Sequence Read Archive (DRA)
Genomic Expression Archive (GEA)
MetaboBank
BioProject
BioSample
Japanese Genotype-phenotype Archive (JGA)
Submission portal D-way

スパコン

NIG SuperComputer