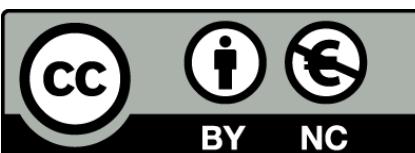


岡山大バイオインフォマティクス ワークショップ

#3-2 RNA-seqデータをSRAに登録する

国立遺伝学研究所 大量遺伝情報研究室 坂本美佳



この講義の目的

- ・ 1/31のオンライン講義@谷澤先生を踏まえて
- ・ どのようにSRA登録作業を進めるか
- ・ 実際に行われた登録作業を題材に説明します

SRA (Sequence Read Archive)

DDBJにあるのはDRA (DDBJ Sequence Read Archive) と呼ばれます

- ・ いわゆる「公共NGSデータ」
- ・ 次世代シークエンサー (NGS) で取ったさまざまなデータを登録・保存
- ・ 今回の実習で使ったRNA-seqデータもSRAに登録されたもの

実習で使用したデータ SRA accessionが記載されているところ

PLOS ONE

RESEARCH ARTICLE

Fucosyltransferase 8 (FUT8) and core fucose expression in oxidative stress response

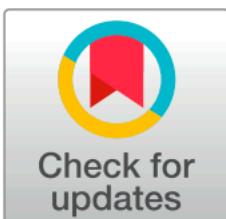
Yuki M. Kyunai¹, Mika Sakamoto², Mayuko Koreishi³, Yoshio Tsujino⁴, Ayano Satoh^{1*}

1 Department of Applied Chemistry and Biotechnology, Faculty of Engineering, Okayama University, Okayama, Japan, **2** National Institute of Genetics, ROIS, Mishima, Shizuoka, Japan, **3** Graduate School of Interdisciplinary Science and Engineering in Health Systems, Okayama University, Okayama, Japan, **4** Graduate School of Science, Technology, and Innovation, Kobe University, Kobe, Hyogo, Japan

* ayano113@cc.okayama-u.ac.jp

Abstract

GlycoMaple is a new tool to predict glycan structures based on the expression levels of 950 genes encoding glycan biosynthesis-related enzymes and proteins using RNA-seq data. The antioxidant response, protecting cells from oxidative stress, has been focused



Copyright: © 2023 Kyunai et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Our RNA-seq data set of 5H4PB treatment is available at the DDBJ Sequence Read Archive (DRA): DRR357080–DRR357084. Another RNA-seq data set is publicly available at the NCBI Sequence Read Archive

PLOS ONE | <https://doi.org/10.1371/journal.pone.0281516> February 13, 2023

PLOS ONE

(SRA). This can be also found @ <https://www.ncbi.nlm.nih.gov/bioproject/897738>. All analysis codes except for the RNA-seq data process pipeline are available at: https://github.com/ayanosatoh/organelle_lab.

Funding: This research was funded by JSPS

Introduction

Glycans consist of complex linkages among galactose, mannose, and N-acetylglucosamine, and exist on the cell surface or freely outside the cell. Because sugar-modifying enzymes, such as glycosyltransferases, are thought to be defined by the expression, for example, in the case of mucin-type O-glycans, Core 1 structures are mainly found in mammary glands, while Core 1 type structures increase in expression of alpha 2,3-sialyltransferase

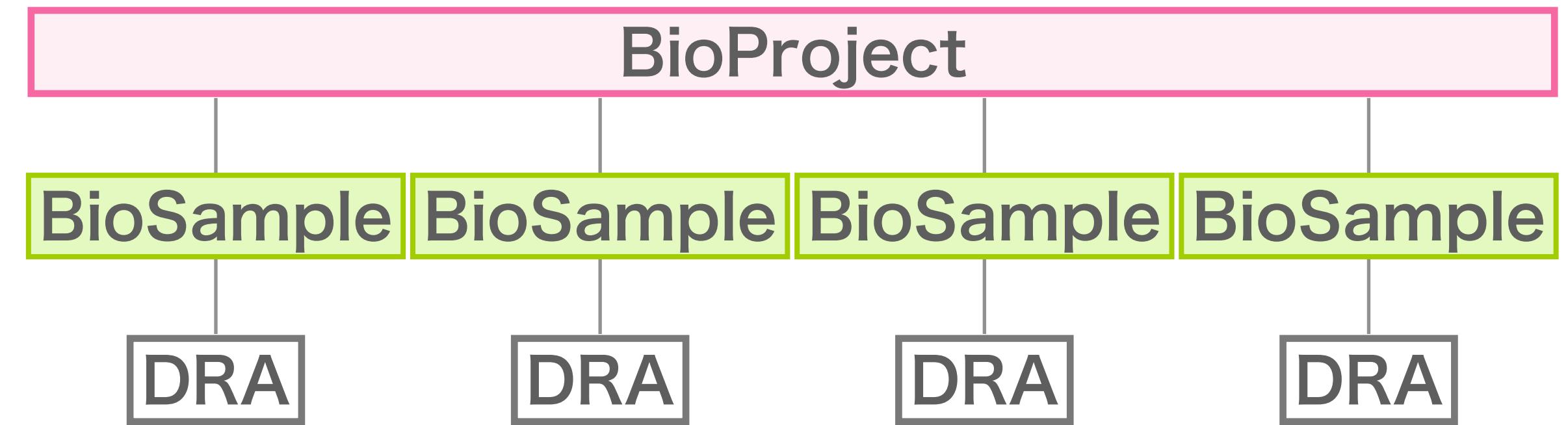
Structures of glycans have been determined by various methods, but these methods have certain limitations. For long sugar chains, it is necessary to digest the sugar chains. Another difficulty is that glycans can consist of different monosaccharides but the same molecular weight; for example, glucose and galactose have the same molecular weight.

データ登録の流れ

DDBJに登録する場合

- BioProject 研究デザインの登録
- BioSample 個別サンプルの登録
- DRA fastqの登録

ロングリードの場合、bamを登録することがある



DDBJのナビゲーションサイト

The screenshot shows the homepage of the DDBJ navigation website at <https://www.ddbj.nig.ac.jp/index.html>. The page features a dark header with the DDBJ logo and navigation links for Services, Supercomputer, Statistics, Activities, and Center Information. A banner at the top right informs about a search keyword mismatch issue and the completion of DDBJ Release 133.0 and DAD Release 103.0. The main content area includes a brief introduction, a grid of service icons (Search, Registration, Services, Supercomputer, Statistics, Activities, and Center Information), and a news sidebar with recent announcements and links.

NEWS

- [復旧](1/24 14:30 ~ 1/25 12:00) D-way一時停止のお知らせ
2024/01/24 メンテナンス BioProject
BioSample DRA GEA DDBJ Center
- DDBJ リリース 133.0, DAD リリース 103.0 完成
2024/01/11 データ公開 DDBJ DDBJ Center
- DDBJ Search キーワード検索の一部不具合について
2023/12/06 お知らせ DDBJ BioProject
BioSample DRA JGA DDBJ Center

[more ▾](#)

SEARCH

DDBJ Search キーワード検索の一部不具合について
DDBJ リリース 133.0, DAD リリース 103.0 完成

ABOUT

生命情報・DDBJセンターは
生命科学研究から生み出されるデータの
共有・解析サービスを提供することで
広く研究活動をサポートしています

Services

- 検索
- 登録
- サービス
- スパコン
- 統計
- 活動
- センターについて

Footer

検索 解析 データベース スパコン
DDBJ Search Vector Screening System Annotated/Assembled Sequences (DDBJ)
getentry WABI (Web API for Biology) Sequence Read Archive (DRA)
NIG SuperComputer

"<https://sc.ddbj.nig.ac.jp/>" を新規タブで開く

<https://www.ddbj.nig.ac.jp/index.html>

DDBJのナビゲーションサイト

The screenshot shows the DDBJ Submission Navigation Site. At the top, there is a banner with the text "DDBJ Search キーワード検索の一部不具合について" and "DDBJ リリース 133.0, DAD リリース 103.0 完成". Below this, the main navigation area is titled "登録ナビゲーション" and includes a breadcrumb trail: ホーム > 登録ナビゲーション.

The registration process consists of two questions:

- Q1. ヒトを対象とした研究データでアクセス制限が必要でしょうか？**
いいえ 変更
- Q2. どのような種類のデータでしょうか？**
全長規模のレプリコン配列(核、オルガネラ、ウイルス、ファージゲノムやプラスミド配列)
メタゲノム配列
転写産物配列
その他の DNA/RNA 配列
NGS による機能ゲノミクスデータ (遺伝子発現、遺伝子制御、エピジェネティクス)
マイクロアレイによる機能ゲノミクスデータ (遺伝子発現、遺伝子制御、エピジェネティクス)
メタボロミクスデータ
プロテオミクスデータ
遺伝学的なバリエントデータ

A red box highlights the option "NGS による機能ゲノミクスデータ (遺伝子発現、遺伝子制御、エピジェネティクス)". A red arrow points from this highlighted text to a callout box containing the following text:

RNA-seqデータを登録するので、
NGSによる機能ゲノミクスデータ（遺伝子発現、遺伝子制御、
エピジェネティクス）を選択

At the bottom of the page, there is a footer section with links to various DDBJ services and a note about opening the URL in a new tab.

DDBJのナビゲーションサイト

The screenshot shows a web browser displaying the DDBJ registration navigation site at ddbj.nig.ac.jp. The page is titled "登録ナビゲーション". A red box highlights the following text:

今回の実施例では、
BioProjectとBioSampleを登録してから、
DRAにfastqを登録します

A red arrow points from this text to the "BioProject" and "BioSample" links in the sidebar under the "概要" section. The sidebar also includes links for "DRA" and "GEA". The main content area contains two questions: Q1 (Hit target research data access restrictions) and Q2 (Data type). Below these is a section titled "データ登録の流れ" (Registration flow) with a numbered list of steps:

1. D-way 登録システムで生リードを DRA に登録。
2. DRA の登録途中もしくは登録前に、D-way 登録システムでプロジェクトを BioProject に、サンプルを BioSample に登録。
3. 解析済みデータを scp/sftp で GEA 登録用ディレクトリにアップロード。
4. D-way 登録システムで GEA のために登録した DRA submission を選択。D-way 登録システムで MAGE-TAB テンプレートファイルにメタデータを記入して登録。

At the bottom of the page, there is a footer with various links categorized by DNA helix icons.

DDBJのナビゲーションサイト

The screenshot shows a web browser window for the DDBJ registration site at ddbj.nig.ac.jp. The page title is "登録ナビゲーション". The main content area contains two questions:

- Q1. ヒトを対象とした研究データでアクセス制限が必要でしょうか？**
いいえ [変更](#)
- Q2. どのような種類のデータでしょうか？**
NGSによる機能ゲノミクスデータ (遺伝子発現, 遺伝子制御, エピジェネティクス) [変更](#)

Below these questions is a section titled "データ登録の流れ" (Flow of data registration) with the following steps:

- 概要**: D-way 登録システムでプロジェクトを BioProject に登録。BioProject で研究のゴールを説明します。
- BioProject**: [前へ](#) [次へ](#)
- BioSample**
- DRA**
- GEA**

At the bottom of the main content area is a button labeled "固定リンクのコピー" (Copy fixed link).

Below the main content is a decorative footer bar featuring DNA helix icons and links to various DDBJ services:

検索	解析	データベース	スバコン
DDBJ Search	Vector Screening System	Annotated/Assembled Sequences (DDBJ)	NIG SuperComputer
getentry	WABI (Web API for Biology)	Sequence Read Archive (DRA)	
ARSA	DDBJ FTP Site	Genomic Expression Archive (GEA)	
TXSearch		MetaboBank	
		BioProject	
		BioSample	
		Japanese Genotype-phenotype Archive (JGA)	
			Submission portal D-way

D-way



今日はアカウントを作らないで
聞いていてください

D-way

The screenshot shows the DDBJ D-way interface at the URL trace.ddbj.nig.ac.jp. The top navigation bar includes links for D-way TOP, BioProject, BioSample, DRA, and GEA. The user account 'nig_msaka' is logged in. A red circle highlights the 'DRA' link in the top menu. Below the menu, a yellow banner displays the text 'Disabilities in the DDBJ Search keyword search functionalities' and 'DDBJ Rel. 133.0, DAD Rel. 103.0 Completed'. The main content area starts with the heading 'Account: nig_msaka'. It contains instructions for submission, mentioning the need to register a public key and providing links for submission navigation and research data from human subjects. Below this, there are three sections: 'BioProject', 'BioSample', and 'DDBJ Sequence Read Archive (DRA)'. Each section provides a brief description and a 'Manual' link. A red box on the right side of the page contains Japanese text: 'DRAを登録するときに BioProjectとBioSampleと一緒に登録できるが、あらかじめアクセスションをとっておいたほうが楽 (な気がする)'.

nig_msaka | Account | Password | Logout

D-way TOP | BioProject | BioSample | DRA | GEA

Disabilities in the DDBJ Search keyword search functionalities

DDBJ Rel. 133.0, DAD Rel. 103.0 Completed

Account: nig_msaka

登録ナビゲーションサイト: データに関するいくつかの質問に応えると登録方法が案内されます。
Submission navigation site: By answering several questions about your data, you will be guided how to submit your data.

Please [register public key](#) at the 'Account' (top right) to submit your data to DRA and GEA.
DRA と GEA ヘデータを登録するためには Account (右上) から [公開鍵を登録する](#)必要があります。

When you are submitting data derived from human subjects, please read "[Submission of research data from human subjects](#)".
ヒトに由来するデータを登録する場合は「[ヒトを対象とした研究データの登録について](#)」を熟読してください。

! Use only ASCII characters (English letters) throughout your submission. Non-ASCII characters (for example, Japanese and special characters) are not accepted.
登録には ASCII 文字 (英語文字) のみを使用してください。非 ASCII 文字 (日本語や特殊文字など) は受け付けていません。

! If there is no reply from submitters after three months of initial contact from us, submissions will be cancelled.
DDBJ センターから登録者に問い合わせた後三か月以上回答が無い場合は Submission をキャンセルいたします。

BioProject

A collection of biological data related to a single initiative, originating from a single organization or from a consortium.
You can register a project here and later submit and link sequence data to the project. [Manual](#)
ここからプロジェクトのみを登録し、後から配列データを登録してプロジェクトに関連付けることができます。 [マニュアル](#)

BioSample

Descriptions of biological source materials used to generate experimental data in any of DDBJ's primary data archives.
You can register sample(s) here and later submit and link sequence data to the sample(s). [Manual](#)
ここからサンプルのみを登録し、後から配列データを登録してサンプルに関連付けることができます。 [マニュアル](#)

DDBJ Sequence Read Archive (DRA)

DRA stores raw and aligned sequence data from next-generation sequencing platforms.
You can complete both your BioProject as well as BioSample submissions within DRA submission interface. [Manual](#)
DRA 登録インターフェースでまとめて BioProject, BioSample と DRA メタデータを登録することができます。 [マニュアル](#)

Genomic Expression Archive (GEA)

High-throughput sequencing and microarray submissions for functional genomics experiments that examine gene expression, regulation, epigenetics or genome variation. [Overview](#)
遺伝子発現、遺伝子発現制御、エピジェネティクスやゲノム変異を解析している機能ゲノミクス実験からのシークエンシングとマイクロアレイデータの登録。 [Overview](#)

DRAを登録するときに
BioProjectとBioSampleと一緒に登録できるが、
あらかじめアクセスションをとっておいたほうが楽
(な気がする)

BioProject

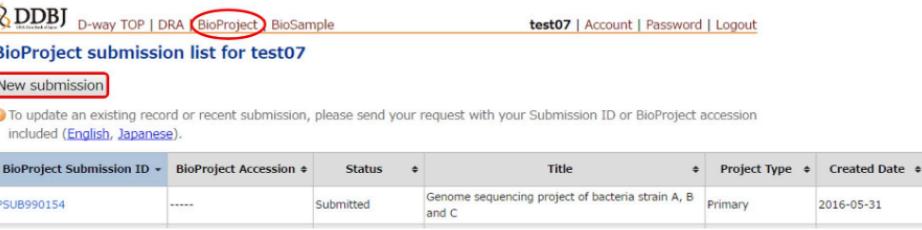
研究・実験デザイン全体

プロジェクトの登録が必要な場合
新規プロジェクトの登録
アクセッション番号
アンブレラプロジェクトの登録
アンブレラプロジェクトへのリンク
ヒトデータの登録
プロジェクトの公開
プロジェクトの更新
論文情報の追加
プロジェクトとデータのリンク
内容に関する問い合わせ

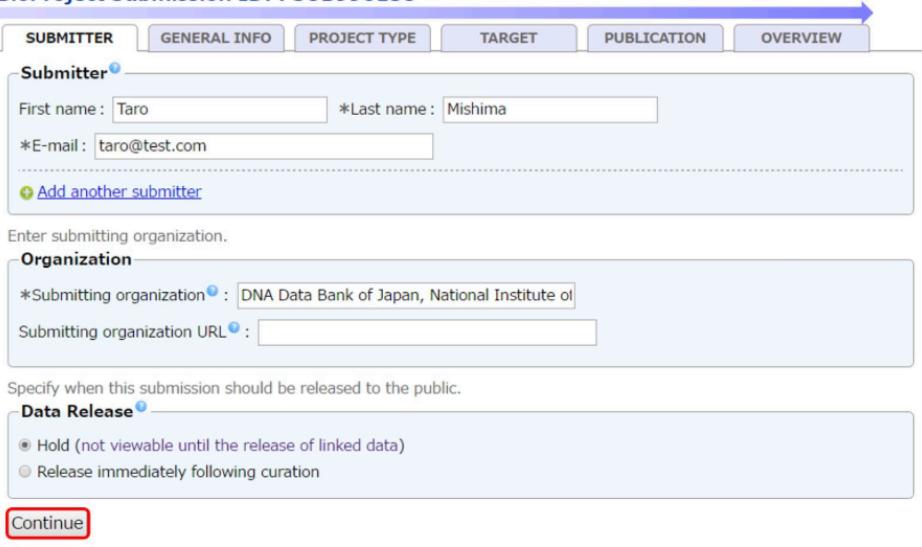
データ登録において BioProject が必要かどうかは「[登録ナビゲーションサイト](#)」でご確認ください。単一のプラスミド、ウイルスやオルガネラゲノムのシーケンスといった1つ(もしくは少数)のアクセッション番号しかリンクされない場合、BioProject 登録は必須ではありません。

新規プロジェクトの登録

登録用アカウントを取得します。
D-way にログインします。ウェブサイト上部にある“BioProject”メニューから BioProject 登録ページに移動します。BioProject ページ内の [New submission] をクリックし、新規プロジェクト登録を作成します。



BioProject の新規登録作成
BioProject を新規登録する場合は左のタブから順番に内容を英語で入力していきます。 [BioProject 入力項目の説明](#)



プロジェクトの入力

機能アノテーションが付されたゲノム配列を [DDBJ](#) に登録する場合、[Locus tag prefix](#) を [BioSample](#) で登録します。2022年11月9日に BioProject の Locus tag prefix 登録口は閉鎖され、prefix 登録は BioSample に一本化されました。

最後の “OVERVIEW” で内容を確認したうえで [Submit] をクリックして登録します。



BioProject

研究・実験デザイン全体

Excelで登録内容をまとめてからD-wayに入力しています

The screenshot shows an Excel spreadsheet titled "BioProject_ayanoRNAseq8-12". The spreadsheet is organized into columns A through H and rows 1 through 14. The data is categorized into three main sections: Submitter, Organization, and Data release.

	A	B	C	D	E	F	G	H
1	metadata1	discription1	mandatory1	content1	content1			
2	Submitter	First name		Ayano	Mika			
3		Last name	必須	Satoh	Sakamoto			
4		E-mail	必須	ayano113@cc.okayama-u.ac.jp	msaka@nig.ac.jp			
5	Organization	Submitting organization	必須	Okayama University	National Institute of Genetics			
6		Submitting organization URL		https://www.okayama-u.ac.jp	https://www.nig.ac.jp/			
7	Data release	Data Release		Hold				
8								
9								
10	BioProject Submission ID: PSUB017111		PRJDB13297					
11								
12								
13								
14								

Key observations from the spreadsheet:

- Submitter:** Rows 2-4. The first row (row 2) has "Submitter" in bold. The "First name" field (B2) contains "Ayano", which is highlighted in yellow. The "Last name" field (C2) contains "Satoh", which is highlighted in red. The "E-mail" field (D2) contains "ayano113@cc.okayama-u.ac.jp", which is highlighted in yellow.
- Organization:** Rows 5-6. The first row (row 5) has "Organization" in bold. The "Submitting organization" field (B5) contains "Okayama University", which is highlighted in yellow. The "Submitting organization URL" field (D5) contains "https://www.okayama-u.ac.jp", which is highlighted in yellow.
- Data release:** Row 7. The first row (row 7) has "Data release" in bold. The "Data Release" field (B7) contains "Hold".
- Submission IDs:** Row 10. The submission IDs "PSUB017111" and "PRJDB13297" are displayed in bold red text at the bottom of the data section.

BioProject

研究・実験デザイン全体

スクリーンショット：BioProject_ayanoRNAseq8-12 の Google Sheets ワークシート

このスクリーンショットは、Google Sheets の操作画面を示しています。ワークシートの名前は「BioProject_ayanoRNAseq8-12」です。

表の構造：

	A	B	C	D	E	F	G	H	I	J	K	L
1	metadata2	discription2	mandatory2	content2	dropdown2							
2	Project Description	Project title	必須	RNA-seq, 5H4PB treated human keratinocyte cell line HaCaT								
3		Description	必須	Human keratinocyte cell line, HaCaT was treated with DMSO (control) or with 100 µM 5H4PB for 24h subjected to RNA extraction for RNA-seq.								
4		Private comments to DDBJ staff,										
5		Relevance		Medical.Industrial	リストから選択							
6		Relevance description	Otherを選択した時必須									
7	Umbrella BioProject	Initiative description	既登録プロジェクトの子である場合必須									
8		Umbrella BioProject accession	既登録プロジェクトの子である場合必須									
9	External Links	Link description										
10		URL										
11	Grants	Agency										
12		Agency abbreviation										
13		Grant ID										
14		Grant title										
15	Consortium	Consortium name										
16		Consortium URL										
17												
18												
19												
20												
21												
22												
23												

注釈：

- セル D5 と E5 が黄色で強調表示されています。
- セル D5 に「Medical.Industrial」が入力されています。
- セル E5 には「リストから選択」というテキストがあります。
- セル D6 と E6 が赤い点線で囲まれています。
- セル D6 に「Otherを選択した時必須」が記載されています。
- セル D7 と E7 が赤い点線で囲まれています。
- セル D7 に「既登録プロジェクトの子である場合必須」が記載されています。
- セル D8 と E8 が赤い点線で囲まれています。
- セル D8 に「既登録プロジェクトの子である場合必須」が記載されています。
- 右側に大きな日本語注釈「この項、不要かも（2024年1月時点）」があります。

BioProject

研究・実験デザイン全体

<https://www.ddbj.nig.ac.jp/bioproject/project-info.html#Project-type>

Screenshot of the BioProject submission interface in Google Sheets, showing the 'Project type' section.

The sheet is titled "BioProject_ayanoRNAseq8-12". The columns are labeled A through H. The rows represent project metadata fields, numbered 1 to 15.

	A	B	C	D	E	F	G	H
1	metadata3	discription3	mandatory3	content3	dropdown3			
2	Project data type	Project data type	必須	Transcriptome or Gene Expression	リストから選択			
3		Project data type description	Otherを選択した時必須					
4	Sample scope	Sample scope	必須	Monoisolate	リストから選択			
5	Material	Material	必須	Transcriptome	リストから選択			
6	Capture	Capture	必須	Whole	リストから選択			
7		Target description	Otherを選択した時必須					
8	Methodology	Methodology	必須	Sequencing				
9		Methodology description	Otherを選択した時必須					
10	Objective	Objective		Raw Sequence Reads	リストから選択			
11	Locus tag prefix	Locus tag prefix						
12								
13								
14								
15								

The 'Project type' row (row 2) has its value "Transcriptome or Gene Expression" highlighted with a red border. The 'Material' row (row 5) has its value "Transcriptome" highlighted with a red border. The 'Capture' row (row 6) has its value "Whole" highlighted with a red border. The 'Methodology' row (row 8) has its value "Sequencing" highlighted with a red border. The 'Objective' row (row 10) has its value "Raw Sequence Reads" highlighted with a red border.

BioProject

研究・実験デザイン全体

項目の詳しい説明は

<https://www.ddbj.nig.ac.jp/bioproject/project-info.html#Project-type>

Project data type	Sample scope	Material	Capture	Methodology	Objective
Genome Sequencing	Monoisolate ✓	Genome	Whole ✓	Sequencing ✓	Raw Sequence Reads ✓
Clone Ends	Multiisolate	Partial Genome	Clone Ends	Array	Sequence
Epigenomics	Multi-species	Transcriptome ✓	Exome	Mass Spectroscopy	Analysis
Exome	Environment	Reagent	Targeted Locus/Loci	Other	Assembly
Map	Synthetic	Proteome	Random Survey		Annotation
Metagenome	Other	Phenotype	Other		Variation
Phenotype and Genotype		Other			Epigenetic Markers
Proteome					Expression
Random Survey					Maps
Targeted Locus (Loci)					Phenotype
Transcriptome or Gene Expression ✓					Other
Variation					
Other					

BioProject

研究・実験デザイン全体

BioProject_ayanoRNAseq8-12

Human skin keratinocyte cell line

D6

Organism information

Organism name	必須	Homo sapiens
Taxonomy ID		9606
Strain, breed, cultivar(A)	AまたはBどちらか	HaCaT
Isolate name or label(B)	AまたはBどちらか	
Description		Human skin keratinocyte cell line

Environmental sample information

Environmental sample name	Sample scope=Environmentの時必須
Environmental sample description	

General Properties

Cellularity	
Reproduction	
Haploid genome size	
Ploidy	

Organism Replicons

Name	
Type	
Location	
Size	
Description	

Phenotype

Disease	
Biotic Relationship	
Trophic Level	

Prokaryote Morphology

Shape	原核生物の場合
Gram	原核生物の場合
Motility	原核生物の場合
Enveloped	原核生物の場合
Endospores	原核生物の場合

Ecological Environment

Habitat	
Salinity	
Oxygen requirement	
Temperature range	
Optimum Temperature	

Publication

PublicationはあとでDDBJに連絡すれば追加できる

Organism informationの項だけでOKでした

BioSample

サンプル個々の情報

<https://www.ddbj.nig.ac.jp/biosample/attribute.html>

The screenshot shows a web browser window for the DDBJ website (www.ddbj.nig.ac.jp). The URL in the address bar is <https://www.ddbj.nig.ac.jp/biosample/attribute.html>. The page title is "BioSample" and the sub-section is "サンプル属性". The navigation menu includes Home, Submission, Sample Attribute, FAQ, Search, and About BioSample. The main content area displays a form for selecting sample attributes under the heading "サンプル属性". A dropdown menu titled "Sample type (Core Package)" lists various categories: Standard (selected), Pathogen, Omics, SARS-CoV-2: clinical or host-associated, SARS-CoV-2: wastewater surveillance, Microbe, Model organism or animal, Metagenome or environmental, Invertebrate, Human, Plant, Virus, Beta-lactamase, and Genome, metagenome or marker sequences (Mlxs compliant). Below this is a "DEFINITION" section with links to "いくつかのパッケージの登録例" and "パッケージ一覧". The top of the page has a banner for "DDBJ Search キーワード検索の一部不具合について" and "DDBJ リリース 133.0, DAD リリース 103.0 完成". The top right corner shows "DDBJ Web Sites", "利用規約", "問合せ", and "English".

BioSample

サンプル個々の情報

- *は必須項目
- 入力できるところはできるだけ記載を
- 不明なところは
not applicable か **missing** とする

詳細は以下を参照してください

<https://www.ddbj.nig.ac.jp/biosample/overview.html>

Human

A template file for attributes: [a template file](#)

Humanパッケージ

Name	Description
<u>sample_name</u> *	sample name は登録者がサンプルに付ける名前です。Submission 単位でユニークな name を付けます。
<u>sample_title</u> *	タイトルはサンプルをよく表す簡潔なものを記入します。タイトルは Submission においてユニークである必要があります。例: 1) Escherichia coli O104:H4 str. C227-11 clinical isolate 2010_333_NC-6; 2) CD8+ T cells from female TSG6-knockout BALB/c mouse; 3) Human metagenome isolated from urine of healthy female.
<u>description</u>	サンプルに対する簡潔な補足情報。
<u>organism</u> *	NCBI Taxonomy database [↗] に登録されている最も下位のランクの生物名 (適切な場合は species まで)。データベースに登録されていない場合、未登録の生物に関する情報をできるだけ記入してください。DDBJ スタッフが NCBI Taxonomy に未登録の生物を申請します。
<u>taxonomy_id</u>	NCBI Taxonomy identifier. 個別の生物、メタゲノムや環境サンプル [↗] に割り当てられています。データベースに登録されていない場合、空欄にします。DDBJ スタッフが NCBI Taxonomy に未登録の生物を申請後、割り当てられた TaxID が自動的に記入されます。
<u>bioproject_id</u>	関連する BioProject アクセッション番号 (PRJDB)
<u>isolate</u> *	サンプルの得られた individual isolate
<u>age</u> *	サンプリング時の年齢。規模は種や研究に依存する。(例: アメーバに対する秒、樹木に対する世紀)
<u>dev_stage</u>	サンプリング時における生物の発生段階
<u>tissue</u> *	サンプルの組織の名称
<u>sex</u> *	サンプル生物の性別
<u>biomaterial_provider</u> *	実験材料の提供元(例: culture collection identifier, Principal Investigator, 研究室名)
<u>sample_type</u>	サンプル種別。例 cell culture, mixed culture, tissue sample, whole organism, single cell, metagenomic assembly, primary cell
<u>collection_date</u> *	サンプルを採取した日付を ISO 8601 形式で記載します。スラッシュによる範囲指定に対応しています。例: 2008-01-23T19:23:10Z, 2008-01-23, 2008-01, 2008, 1952-10-21T11:43Z/1952-10-21T17:43Z, 1952-10-21/1953-02-15, 1952-10/1953-02, 1952/1953。日時は国際標準時 (UTC) で記載します。タイムゾーンの指定が無い時間は UTC として処理され、UTC ではない時間は UTC に変換されます。
<u>geo_loc_name</u> *	サンプルを得た地域を国、または、海洋の名称で示し、続けて地方・地域を示します。国名は <u>country_list</u> の名前を選択し、コロンに続けて地方・地域を記載します。一つの属性内で複数の地点を記載することは禁止しています。例 Japan:Kanagawa, Hakone, Lake Ashi
<u>cell_line</u>	cell line の名称
<u>cell_subtype</u>	Cell subtype
<u>cell_type</u>	サンプルの細胞のタイプ
<u>culture_collection</u>	培養細胞の保管施設と ID (例: ATCC:26370) 生きている微生物やウイルスの培養系、および細胞株を記載するのに用います。書式は <u>institute</u> のリスト [↗] を参照してください。
<u>disease</u>	関連する病気の名称
<u>disease_stage</u>	サンプリング時における病気の段階
<u>ethnicity</u>	Ethnicity of the subject
<u>health_state</u>	収集時におけるサンプルの健康あるいは病気の状態
<u>karyotype</u>	Karyotype
<u>phenotype</u>	サンプル生物の表現型。Phenotypic quality Ontology (PATO) (v1.269) の用語は このリンク [↗] を参照してください。
<u>population</u>	For human: ethnicity or population of origin; for plants: filial generation, number of progeny, genetic structure
<u>race</u>	Race
<u>treatment</u>	Treatment, treatment protocol

<https://www.ddbj.nig.ac.jp/biosample/attribute.html>

BioSample

サンプル個々の情報

- *は必須項目
- 入力できるところはできるだけ記載を
- 不明なところは
not applicable か **missing** とする

詳細は以下を参照してください

<https://www.ddbj.nig.ac.jp/biosample/overview.html>

このデータセットはHumanパッケージで登録したが、
Omicsパッケージの方が良かったかも。。。

Name	Description	Omicsパッケージ
<u>sample_name</u> *	sample name は登録者がサンプルに付ける名前です。Submission 単位でユニークな name を付けます。	
<u>sample_title</u> *	タイトルはサンプルをよく表す簡潔なものを記入します。 タイトルは Submission においてユニークである必要があります。 例: 1) Escherichia coli O104:H4 str. C227-11 clinical isolate 2010_333_NC-6; 2) CD8+ T cells from female TSG6-knockout BALB/c mouse; 3) Human metagenome isolated from urine of healthy female.	
<u>description</u>	サンプルに対する簡潔な補足情報。	
<u>organism</u> *	NCBI Taxonomy database [2] に登録されている最も下位のランクの生物名 (適切な場合は species まで)。データベースに登録されていない場合、未登録の生物に関する情報をできるだけ記入してください。 DDBJ スタッフが NCBI Taxonomy に未登録の生物を申請します。	
<u>taxonomy_id</u>	NCBI Taxonomy identifier. 個別の生物、メタゲノムや環境サンプル [2] に割り当てられています。データベースに登録されていない場合、空欄にします。 DDBJ スタッフが NCBI Taxonomy に未登録の生物を申請後、割り当てられた TaxID が自動的に記入されます。	
<u>bioproject_id</u>	関連する BioProject アクセッション番号 (PRJDB)	
<u>strain</u>	微生物や真核生物の株名。	
<u>isolate</u>	サンプルの得られた individual isolate	
<u>breed</u>	品種の名称。主に家畜化された動物に用いられます。	
<u>cultivar</u>	植物の栽培品種名。	
<u>ecotype</u>	遺伝学的に生育環境への適応を反映した表現型特性を示す種内集団(例: Columbia)	
<u>isolation_source</u>	サンプルが得られた生物学的サンプルに関する、物理的、環境的、かつまたは、地理的な由来	
<u>age</u>	サンプリング時の年齢。規模は種や研究に依存する。(例: アメーバに対する秒、樹木に対する世紀)	
<u>dev_stage</u>	サンプリング時における生物の発生段階	
<u>tissue</u>	サンプルの組織の名称	
<u>sex</u>	サンプル生物の性別	
<u>biomaterial_provider</u>	実験材料の提供元(例: culture collection identifier, Principal Investigator, 研究室名)	
<u>sample_type</u>	サンプル種別。例 cell culture, mixed culture, tissue sample, whole organism, single cell, metagenomic assembly, primary cell	
<u>collection_date</u> *	サンプルを採取した日付を ISO 8601 形式で記載します。スラッシュによる範囲指定に対応しています。例: 2008-01-23T19:23:10Z, 2008-01-23, 2008-01, 2008, 1952-10-21T11:43Z/1952-10-21T17:43Z, 1952-10-21/1953-02-15, 1952-10/1953-02, 1952/1953。日時は国際標準時 (UTC) で記載します。タイムゾーンの指定が無い時間は UTC として処理され、UTC ではない時間は UTC に変換されます。	
<u>geo_loc_name</u> *	サンプルを得た地域を国、または、海洋の名称で示し、続けて地方・地域を示します。国名は country_list の名前を選択し、コロンに続けて地方・地域を記載します。一つの属性内で複数の地点を記載することは禁止しています。例 Japan:Kanagawa, Hakone, Lake Ashi	
<u>lat_lon</u>	サンプルが採取された位置の地理的座標。(書式 : d[d.ddddddd] d[dd.ddddddd]) (例: "47.94 N 28.12 W", "45.0123 S 4.1234 E") 小数点以下の数字は分秒ではなく小数で記載してください。位置情報が無い、あるいは、記載することが不適切な場合は "missing" と記入してください。	
<u>biological_replicate</u>	Biological replicate	
<u>antibody</u>	Antibody name, provider name, lot number, if used.	
<u>cell_line</u>	cell line の名称	
<u>cell_type</u>	サンプルの細胞のタイプ	
<u>chem_administration</u>	List of chemical compounds administered to the host or site where sampling occurred, and when (e.g. antibiotics, N fertilizer, air filter); can include multiple compounds. For Chemical Entities of Biological Interest ontology (CHEBI) (v1.72), please see this list [2] .	
<u>culture_collection</u>	培養細胞の保管施設と ID (例: ATCC:26370) 生きている微生物やウイルスの培養系、および細胞株を記載するのに用います。書式は institute のリスト [2] を参照してください。	
<u>disease</u>	関連する病気の名称	
<u>disease_stage</u>	サンプリング時における病気の段階	
<u>genetic_modification</u>	Genetic modification	
<u>genotype</u>	Observed genotype	
<u>growth_protocol</u>	Free-text growth protocol	
<u>infection</u>	Infection	
<u>karyotype</u>	Karyotype	
<u>passage_history</u>	Number of passages and passage method	
<u>phenotype</u>	サンプル生物の表現型。Phenotypic quality Ontology (PATO) (v1.269) の用語はこのリンク [2] を参照してください。	
<u>specimen_voucher</u>	標本(動植物個体の一部 または 全体)が維持管理されている管理団体と ID。(書式 : [<institution_code>:<collection_code>:]<specimen_id>) (例 : UAM:Mamm:52179) <collection_code> が存在しない場合は記載不要です。<institution_code> は こちら [2] を参照してください。	
<u>stress</u>	Stress	
<u>temp</u>	サンプリング時のサンプルの温度	
<u>time</u>	Time	
<u>treatment</u>	Treatment, treatment protocol	

<https://www.ddbj.nig.ac.jp/biosample/attribute.html>

BioSample

サンプル個々の情報

D22 x ✓ fx

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	metadata1	discription1	mandatory1	content1	content2								
2	Submitter	First name	必須	Ayano	Mika								
3		Last name	必須	Satoh	Sakamoto								
4		E-mail	必須	ayano113@cc.okayama-u.ac.jp	msaka@nig.ac.jp								
5	Organization	Organization		Okayama University	National Institute of Genetics								
6		Submitting organization	必須	Okayama University									
7		Submitting organization URL		https://www.okayama-u.ac.jp									
8	Data release	Data Release		Hold	リストから選択								
9													
10	BioSample Submission ID: SSUB021383												
11													
12													

Submitter General info Sample type Human + 準備完了 100%

A1 x ✓ fx metadata2

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	metadata2	discription2	mandatory2	content2	dropdown2										
2	External Links	External links													
3		Link description													
4		URL													
5															
6															
7															
8															
9															
10															
11															
12															
13															

Submitter General info Sample type Human + 準備完了 100%

BioSample

サンプル個々の情報

Humanパッケージで登録しました

	A	B	C	D
1	metadata3	description3	mandatory3	content3
2	Core Package	Genome, metagenome or marker sequences (MIxS compliant)		
3		Functional genomics samples (e.g. transcriptome, epigenetics etc)		
4		Other samples (e.g. transcriptome, epigenetics etc)		Other samples (e.g. transcriptome, epigenetics etc)
5	MIxS	(Meta)Genomic Sequences Sample (MIMS)		
6		Genomic Sequences Sample (MIGS)		
7		Marker Sequences Sample (MIMARKS)		
8	Environmental package	Environmental package (MIxS Sample)		
9				
10				

Submitter | General info | **Sample type** Human | + | 準備完了

	A	B	C	D	E	F	G	H	I	J	K	L
1	*sample_name	*sample_title	description	*organism	taxonomy_id	bioproject_id	*age	*biomaterial_provider	*isolate	*sex	*tissue	cell
2	sample_name*	sample_title*	description	organism*	taxonomy_id	bioproject_id	age*	biomaterial_provider*	isolate*	sex*	tissue*	cell
3	DNBSEQ_OS8	Human keratinocyte cell line HaCaT, treated with DMSO, biological replicate 1		Homo sapien	9606	PRJDB13297	not applicable	not applicable	not applicabl	not applicabl	not applicable	Ha
4	DNBSEQ_OS9	Human keratinocyte cell line HaCaT, treated with DMSO, biological replicate 2		Homo sapien	9606	PRJDB13297	not applicable	not applicable	not applicabl	not applicabl	not applicable	Ha
5	DNBSEQ_OS10	Human keratinocyte cell line HaCaT, treated with 5H4PB, biological replicate 1		Homo sapien	9606	PRJDB13297	not applicable	not applicable	not applicabl	not applicabl	not applicable	Ha
6	DNBSEQ_OS11	Human keratinocyte cell line HaCaT, treated with 5H4PB, biological replicate 2		Homo sapien	9606	PRJDB13297	not applicable	not applicable	not applicabl	not applicabl	not applicable	Ha
7	DNBSEQ_OS12	Human keratinocyte cell line HaCaT, treated with 5H4PB, biological replicate 3		Homo sapien	9606	PRJDB13297	not applicable	not applicable	not applicabl	not applicabl	not applicable	Ha
8												
9												
10												

Submitter | General info | **Sample type** Human | + | 準備完了

	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	cell_line	cell_subtype	cell_type	culture_colle	dev_stage	disease	disease_stag	ethnicity	health_state	karyotype	phenotype	population	race	sample_type	treatment	replicate	
2	cell_line	cell_subtype	cell_type	culture_colle	dev_stage	disease	disease_stag	ethnicity	health_state	karyotype	phenotype	population	race	sample_type	treatment		
3	HaCaT															biological replicate 1	
4	HaCaT															biological replicate 2	
5	HaCaT															biological replicate 1	
6	HaCaT															biological replicate 2	
7	HaCaT															biological replicate 3	
8																	
9																	
10																	

Submitter | General info | **Sample type** Human | + | 準備完了

DRA

NGSの生データ(fastq)

- 準備

- ✓ D-wayで new submission > submission IDを取得

fastqがいくつあっても1回でまとめて登録できる
BioProjectが異なっても1回で登録できる
(実際には、異なるBioProjectは分けることが多い)

登録するfastqファイルのあるディレクトリで以下のコマンドを実行

```
md5sum filename
```

("filename"は各fastqファイル名)

- ✓ D-wayでfastqをupload (submission 画面にコマンド表示あり)

fastqが遺伝研スパコンにあるときは、スパコンからscpすると爆速 (ただしスパコンの公開鍵が必要)

DRA

NGSの生データ(fastq)

<https://github.com/ddbj/submission-excel2xml>

The screenshot shows the GitHub repository page for 'submission-excel2xml'. The repository is public and has 90 commits. The commit history lists various changes, including file renames, documentation updates, and bug fixes. A red circle highlights the file 'metadata_dra.xlsx' in the commit list.

komstat bp cell restored 9a436b3 · last month 90 Commits

- .github/workflows Add tests 2 months ago
- bin Rename Excelxml to SubmissionExcel2xml 2 months ago
- example bp cell restored last month
- exe v2.4 last month
- lib Documentation 2 months ago
- spec Remove spec 2 months ago
- .dockerignore Documentation 2 months ago
- .gitignore Gemify 2 months ago
- .rspec Gemify 2 months ago
- Dockerfile Rename Excelxml to SubmissionExcel2xml 2 months ago
- Gemfile Add rake as dependency 2 months ago
- Gemfile.lock Documentation 2 months ago
- JGA_metadata.xlsx template excel path revised last month
- LICENSE.txt Gemify 2 months ago
- README.md v2.4 last month
- Rakefile Abolish automatic xsd dowloads 2 months ago
- Singularity Order sections 2 months ago
- metadata_dra.xlsx bp cell restored last month
- submission-excel2xml.gemspec Documentation 2 months ago

About

Tools for XML submission

ddbj-curators

- Readme
- Apache-2.0 license
- Activity
- Custom properties
- 5 stars
- 9 watching
- 3 forks

Report repository

Releases 3

center name changes Latest on Dec 21, 2023 + 2 releases

Packages

No packages published

Contributors 4

- komstat Kodama Yuichi
- ursm Keita Urashima
- yookuda
- machupicchubeta machupicchubeta

Languages

metadata_dra.xlsxを記入



一括登録用xmlファイルに変換

DRA

NGSの生データ(fastq)

	A	B	C	D
1	Center Name	Lab Name	Hold Until	BioProject accession
2	NIG	Genome Informatics Lab	2023-04-01	PRJDB13297
3				
4	Submitter Name	Submitter E-mail		
5	Ayano Satoh	ayano113@cc.okayama-u.ac.jp		
6	Mika Sakamoto	msaka@nig.ac.jp		
7				
8				
9				

Readme Submission Experiment Run Run-file Admin + 準備完了

デフォルトでつけられるTitle（サンプル条件がわかりやすい名称に変えてOK） Oligo-dTと登録されている場合も多い

	Alias	Title	BioSample Used	Library Name	Library Source	Library Selection	Library Strategy	Library Construction Proto
2	Experiment-1	DNBSEQ-G400 paired end sequencing of SAMD00451567	SAMD00451567	DNBSEQ_OS8	TRANSCRIPTOMIC	PolyA	ssRNA-seq	KAPA mRNA Capture Ki
3	Experiment-2	DNBSEQ-G400 paired end sequencing of SAMD00451568	SAMD00451568	DNBSEQ_OS9	TRANSCRIPTOMIC	PolyA	ssRNA-seq	KAPA mRNA Capture Ki
4	Experiment-3	DNBSEQ-G400 paired end sequencing of SAMD00451569	SAMD00451569	DNBSEQ_OS10	TRANSCRIPTOMIC	PolyA	ssRNA-seq	KAPA mRNA Capture Ki
5	Experiment-4	DNBSEQ-G400 paired end sequencing of SAMD00451570	SAMD00451570	DNBSEQ_OS11	TRANSCRIPTOMIC	PolyA	ssRNA-seq	KAPA mRNA Capture Ki
6	Experiment-5	DNBSEQ-G400 paired end sequencing of SAMD00451571	SAMD00451571	DNBSEQ_OS12	TRANSCRIPTOMIC	PolyA	ssRNA-seq	KAPA mRNA Capture Ki

Readme Submission Experiment Run Run-file Admin + 準備完了 130%

	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Library Construction Protocol		Instrument	Spot Type	Nominal Length	Nominal Sdev	Spot Length							
2	KAPA mRNA Capture Kit, MGIEasy RNA Directional Library Prep Set, DNBSEQ-G400RS	DNBSEQ-G400	paired (FR)	250	300									
3	KAPA mRNA Capture Kit, MGIEasy RNA Directional Library Prep Set, DNBSEQ-G400RS	DNBSEQ-G400	paired (FR)	250	300									
4	KAPA mRNA Capture Kit, MGIEasy RNA Directional Library Prep Set, DNBSEQ-G400RS	DNBSEQ-G400	paired (FR)	250	300									
5	KAPA mRNA Capture Kit, MGIEasy RNA Directional Library Prep Set, DNBSEQ-G400RS	DNBSEQ-G400	paired (FR)	250	300									
6	KAPA mRNA Capture Kit, MGIEasy RNA Directional Library Prep Set, DNBSEQ-G400RS	DNBSEQ-G400	paired (FR)	250	300									

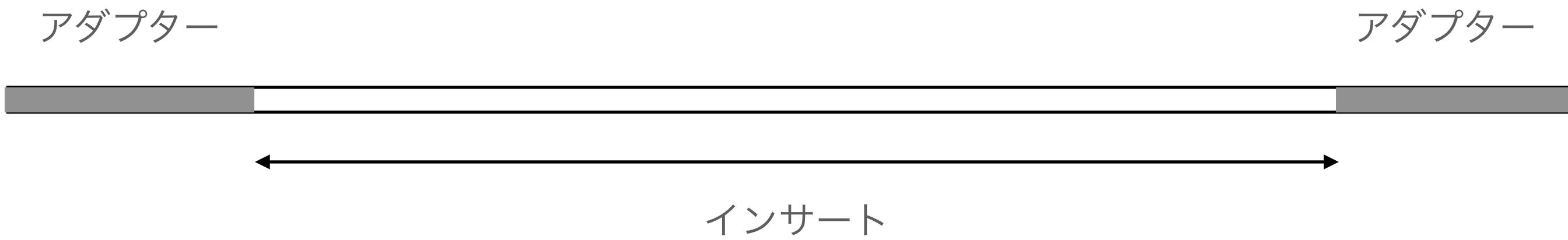
Readme Submission Experiment Run Run-file Admin + 準備完了 130%

insert_sizeに変更
(2024年1月現在)

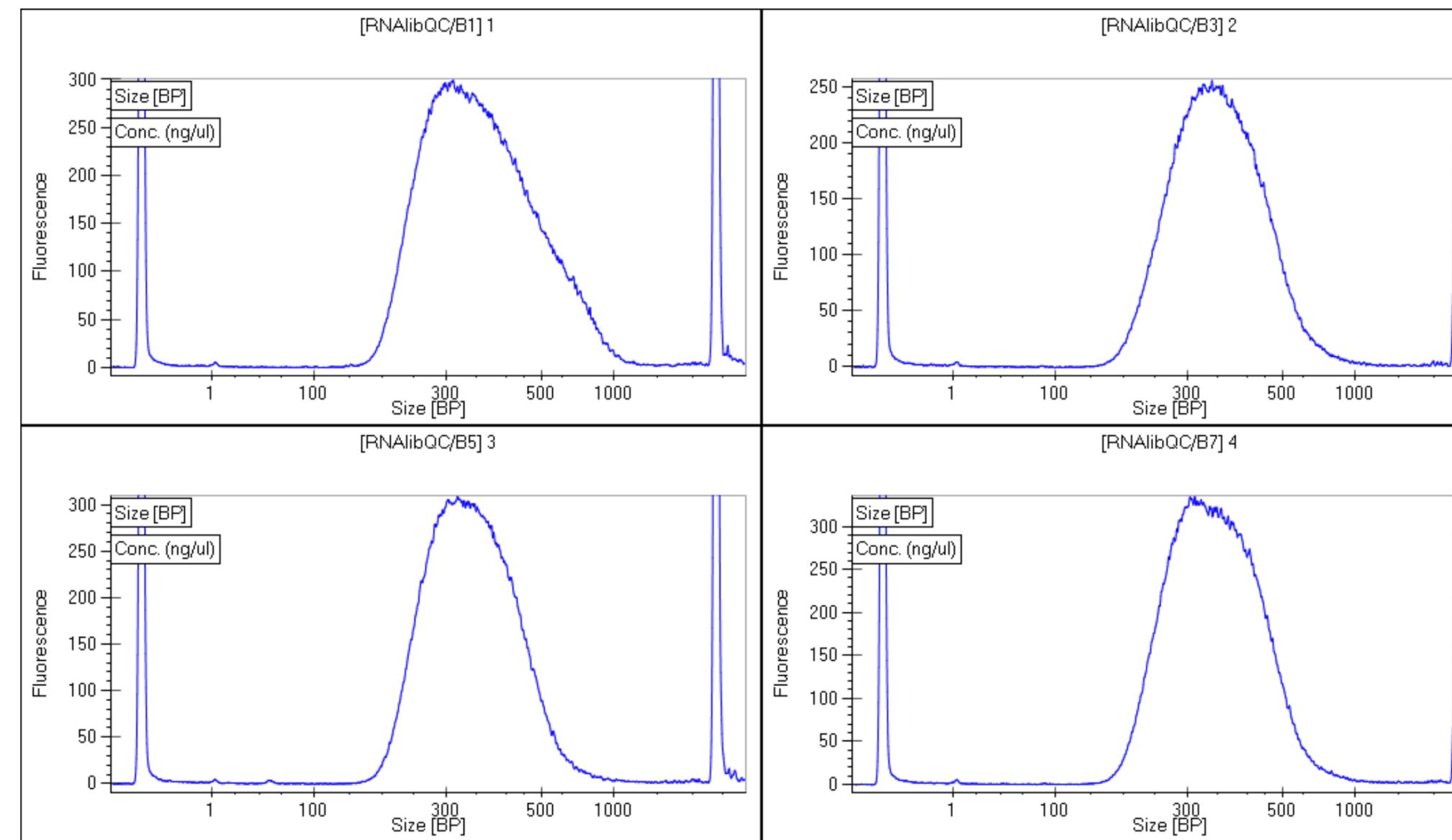
この項目なくなった

DRA

インサートサイズについて



ライブラリの電気泳動像



作業報告書では
「インサートサイズ250baseになるように断片化」
ピークは >300 base にあるが、ライブラリにはアダプター
配列が含まれているので妥当な値

DRA

NGSの生データ(fastq)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Alias	Title	Experiment Referenced															
2	Run-1	DNBSEQ-G400 paired end sequencing of SAMD00451567	Experiment-1															
3	Run-2	DNBSEQ-G400 paired end sequencing of SAMD00451568	Experiment-2															
4	Run-3	DNBSEQ-G400 paired end sequencing of SAMD00451569	Experiment-3															
5	Run-4	DNBSEQ-G400 paired end sequencing of SAMD00451570	Experiment-4															
6	Run-5	DNBSEQ-G400 paired end sequencing of SAMD00451571	Experiment-5															
7																		
8																		

B11 × ✓ f_x Run-5 あらかじめ計算しておいたmd5sumを記入

	A	B	C	G	H	I	J	K	L	M
1	File Name	Run contains files	File Type	MD5 Checksum						
2	DNBSEQ_OS8_Read1.fq.gz	Run-1	generic_fastq	9954b3cc4694fb3a4dbeac54510bc380						
3	DNBSEQ_OS8_Read2.fq.gz	Run-1	generic_fastq	020e17dfcd43e24cefefaf70bca2cd512						
4	DNBSEQ_OS9_Read1.fq.gz	Run-2	generic_fastq	fd9f5c327997aeb96590863e56965bbc						
5	DNBSEQ_OS9_Read2.fq.gz	Run-2	generic_fastq	323302f04f5581f152e8290f79cb4e9d						
6	DNBSEQ_OS10_Read1.fq.gz	Run-3	generic_fastq	2282e23a7c08faa314d3fa50191a3fc1						
7	DNBSEQ_OS10_Read2.fq.gz	Run-3	generic_fastq	7ba1e778a121d1e36620d968a00a4034						
8	DNBSEQ_OS11_Read1.fq.gz	Run-4	generic_fastq	20d493994cbb06e98a5ddc3b3d6b925c						
9	DNBSEQ_OS11_Read2.fq.gz	Run-4	generic_fastq	f8e256b2708e9c4dc1c4a5813a107e4a						
10	DNBSEQ_OS12_Read1.fq.gz	Run-5	generic_fastq	473aba16c924ee9480d8fb7db4a174ce						
11	DNBSEQ_OS12_Read2.fq.gz	Run-5	generic_fastq	0a6965f2647a56b0017145a2f2842c1e						
12										

ここまで記入できたら、submission-excel2xmlでxmlファイルに変換する

DRA

メタデータファイルの変換

- 遺伝研スパコンを使う場合

例・submission id = okym-0001, BioProject = PRJDB99999

```
# singularityコンテナのダウンロード
wget https://ddbj.nig.ac.jp/public/software/submission-excel2xml/excel2xml.simg
singularity exec excel2xml.simg excel2xml_dra -a okym -i 0001 -p PRJDB99999 okym-0001_dra_metadata.xlsx
```

以下のファイルができるので、これらをDRAのメタデータとしてuploadする

- okym-0001_dra_Submission.xml
- okym-0001_dra_Experiment.xml
- okym-0001_dra_Run.xml

DRA

Excelファイルの変換

- Dockerを使う場合（自分のパソコンで）

例・submission id = okym-0001, BioProject = PRJDB99999

```
# Dockerコンテナのビルド
git clone https://github.com/ddbj/submission-excel2xml.git
cd submission-excel2xml
sudo docker build -t excel2xml .

# submission id = okym-0001, BioProject = PRJDB99999 の場合
sudo docker run -v .:/data -w /data excel2xml excel2xml_dra -a okym -i 0001 -p PRJDB99999 okym-0001_dra_metadata.xlsx
```

以下のファイルができるので、これらをDRAのメタデータとしてuploadする

- okym-0001_dra_Submission.xml
- okym-0001_dra_Experiment.xml
- okym-0001_dra_Run.xml

登録完了まで Submitボタンを押してから...

- BioProject, BioSample 即日（ただし混み具合による）
- DRA 数日（ただし混み具合による）

今までの経験から、
メタデータを揃え始めてから数日~1週間でExcelが埋まり、
DRA登録完了までさらに数日
(さらにfastqのアップロード時間がかかる)

慣れれば簡単