

岡山大バイオインフォマティクス ワークショップ

#1~2 RNA-seqハンズオン・前処理とリードカウント

国立遺伝学研究所 大量遺伝情報研究室 坂本美佳

スライドの内容

- ・ 講習の目的と参考書
- ・ 計算資源について
- ・ CUIに慣れましょう
- ・ 遺伝研スパコンの使い方
- ・ RNA-seq解析パイプラインの説明
- ・ 解析実行はhandout1 とhandout2で

自己紹介

略歴

2019年4月 - 現在	国立遺伝学研究所 情報研究系 大量遺伝情報研究室 特任研究員	
2017年7月 - 2019年3月	国立成育医療研究センター研究所 メディカルゲノムセンター	
2015年4月 - 2017年6月	(株) 日本バイオデータ	バイオインフォマティシャンとして働きながら博士課程進学
2011年10月 - 2015年3月	中央大学 理工学部生命科学科（原山研究室） 技術員	大学院再入学
2010年9月 - 2011年9月	日本大学 医学部整形外科学系 臨時職員（実験助手）	
2008年4月 - 2010年8月	埼玉大学大学院 理工学研究科 技術補佐員	
1992年4月 - 1997年7月	埼玉県警察本部 刑事部科学捜査研究所 法医科	

主な仕事

遺伝子アノテーション「遺伝子探し」
アノテーションデータベースの構築

Genome browserToolsDownload

Search

Cats-I: Cats' genome Informatics

Genome database for *Felis catus*

This web site contains genomic data obtained from the study for "AnAms1.0: A high-quality chromosome-scale assembly of a domestic cat *Felis catus* of American Shorthair breed" (Isobe, Matsumoto, Chung, Sakamoto, et al. BioRxiv, 2020. [🔗](#))



AnAms Felis

Keyword Search

Q

e.g. kinase PF01018 "GMP synthase" AnAmsA1_00100
OR: topoisomerase gyrase (default behavior) AND: +"elongation factor" +transcription (Add + to each keyword)
NOT: "elongation factor" -transcription (Add - to exclude from search) Prefix search: ribosom* (ribosome, ribosomal, etc.)

Our cat



Name
Senzu

Breed
American Shorthair

Sex
Female

Genome specs.
20 sequences (19 chromosomes + 1 unplaced)
2.49 Gbp in total

Assembly and annotation version

The current version of the genome assembly is ver1.0 (AnAms1.0).
The genome sequences are also available from the INSDC under the accession number BioProject:PRJDB9879.

The current version of the annotation is ver1.0 revision 1.0.2 (AnAms1.0r1.0.2).

Citation

The data provided in this web site is freely available for academic purposes. Please cite the preprint posted on BioRxiv if you use the data obtained from the web site. DOI: [10.1101/2020.05.19.103788](#)

Inquiries and feedback

Genome Informatics Laboratory, National Institute of Genetics. yn @ nig.ac.jp

Genome browserToolsDownload

Search

About

Red Perilla annotation database

Genome database for *Perilla frutescens*

This web site contains genomic data obtained from the study for "A highly contiguous genome assembly of red perilla (*Perilla frutescens*) domesticated in Japan" (Tamura, et al., DNA research, 2022, 30, 1-8. [🔗](#))


Pfru_yukari_1

Keyword Search

Q

e.g. kinase PF01018 "GMP synthase"
OR: topoisomerase gyrase (default behavior) AND: +"elongation factor" +transcription (Add + to each keyword)
NOT: "elongation factor" -transcription (Add - to exclude from search) Prefix search: ribosom* (ribosome, ribosomal, etc.)

Our resources



Cultivar
Hoko-3

Genome specs.
20 pseudochromosome
1.25 Gbp in total

Assembly and annotation version

The current version of the genome assembly is ver1.0 (Pfru_yukari_1.0).
The genome sequences are also available from the INSDC under the accession number BioProject:PRJDB14288.

The current version of the annotation is ver1.0(Pfru_yukari_1.0).

Citation

The data provided in this web site is freely available for academic purposes. Please cite the following article if you use the data obtained from the web site. DOI: [10.1093/dnares/dsac044](#)
The annotation data also provided on figshare. DOI: [10.6084/m9.figshare.20780995.v2](#)

Inquiries and feedback

Genome Informatics Laboratory, National Institute of Genetics. yn @ nig.ac.jp

この講習の目的

- RNA-seq（バルクRNA-seq）解析を自分の手でできるようにする

✓ fastqの前処理

遺伝研スパコン

✓ カウントデータの取得

✓ 発現変動遺伝子を見つける

ウェブツール

✓ データの可視化（PCA、ヒートマップetc.）

参考書



実験医学別冊

改訂版RNA-Seqデータ解析 WETラボのための超鉄板レシピ

人から非モデル生物まで 公共データの活用も充実

坊農秀雅・編（羊土社）

2023年10月に出版されたばかりです

- 講習の目的と参考書
- **計算環境について**
- CUIに慣れましょう
- 遺伝研スパコンの使い方
- RNA-seq解析パイプラインの説明

計算環境について

スパコン利用？自分のパソコン？

今日の計算環境

1. 遺伝研スパコン
2. 遺伝研スパコン以外の計算サーバ（他大学のスパコン含む）
3. 自分のPC

計算環境について

遺伝研スパコン利用のひと

スパコンにsshでログイン

1. できる
2. できない

計算環境について

スパコン利用？自分のパソコン？

今日使うパソコンのOS

1. Mac
2. Windows
3. その他（Linuxなど）

計算環境について

スパコン利用？自分のパソコン？

使えるストレージ（空きストレージ）

1. 1TB以上
2. 500GB~1TB
3. 256GB~500GB
4. 256GB未満

計算環境について

スパコン利用？自分のパソコン？

- ・ スパコンや高スペックの計算サーバを使うことをおすすめ
数時間から一晩回しっぱなしになります
- ・ メモリは盛れるだけ盛る
- ・ ストレージ（HDDやSSD） 容量も多めに 巨大な中間ファイルが生成されます
- ・ Linux-likeな操作環境（Windowsなら**WSL**利用）
PowerShellのコマンドはLinux（例えばbash）のコマンドとはちがいます

シークエンスデータの管理

巨大データの管理

- NGSのデータ（fastq、bam...）は巨大 この講習で使うデータセットは約40GB、
中間ファイル含めて最終的に150GB程度になります
- 遺伝研スパコンのストレージは1TB/1アカウント 2024年1月現在、大規模利用制度の新規申請はできません
- 必要なデータだけをupload
- 終わったら外付けHDDにバックアップ→スパコンから削除

データバックアップに使えるコマンドやツール

- scp
- rsync
- Aspera

途中で止まっても再開できる、つないでしまえば
あとは放置でOK

シークエンスデータの管理

ファイル名の付け方（あくまでも私見）

- ・ 連番にしておく と スクリプトで処理しやすい リスト（配列）にしてindexで指定できるので、
シェルスクリプトに慣れてくればあまり気にしなくていいかも
- ・ ファイル名-サンプル名-実験条件の対応表を作っておく
- ・ SRAのデータを使うとき → BioSampleに実験条件などの情報がある

諸注意

この講習では**スパコン利用を前提**として内容を組み立てています

- 待ち時間多いです **スパコンは結構混んでいます** **すぐにジョブが入らないことも**
- 講習時間内に終わらないかも **1ファイルの処理に2~3時間かかるステップがあります**
- 何度もやり直すかもしれません **コマンドのタイプミスや、コピペで全角スペースが入ってしまったり**
- 想定外のこともたくさん起こります
- 講師はWindowsをよく知りません **WindowsのひとはWSLを使いましょう**
- 結果の解釈は各自で **見たいもの、興味あるところはそれぞれ・・・**

解析で困ったら

- ・わからなくなったらぐぐりましょう 同じことで困っている人は世界のどこかにいる
- ・できる人に聞きましょう（すぐ近所にいなくてもSNSを活用）
- ・生成AI（ChatGPTやCopilot）に聞くのもいいですね

質問するときの準備

- ログファイル
- エラーメッセージ
- コマンド（特にオプション）
- ツールのバージョン

- 講習の目的と参考書
- 計算環境について
- **CUIに慣れましょう**
- 遺伝研スパコンの使い方
- RNA-seq解析パイプラインの説明

CUIに慣れましょう

CUI (Character User Interface、いわゆるコマンドライン) を

1. 日常的に使っている
2. 使ったことはある
3. 使ったことがない、ターミナルappを開くのも初めて

CUIに慣れましょう

Linux基本コマンド

`ls` 今いる場所のファイルリストを出力

`-l` パーミッションなどの情報も出力

`-a` 不可視ファイルも出力

`cd` ディレクトリの移動

`mkdir` ディレクトリ作成

`less`(または`more`) ファイルの中身を見る

`ls -l`

コマンドのあとに半角スペースをいれてオプションをつけます

`ls -la`

見え方をくらべてみましょう

CUIに慣れましょう

ファイルのパーミッション

ls -l で出力される

d	rwx	rwx	rwx	r:readable	4
				w:writable	2
				x:executable	1
				-:	0
↑ ディレクトリ	_____ オーナー (ファイルの作成者)	_____ グループ	_____ その他		

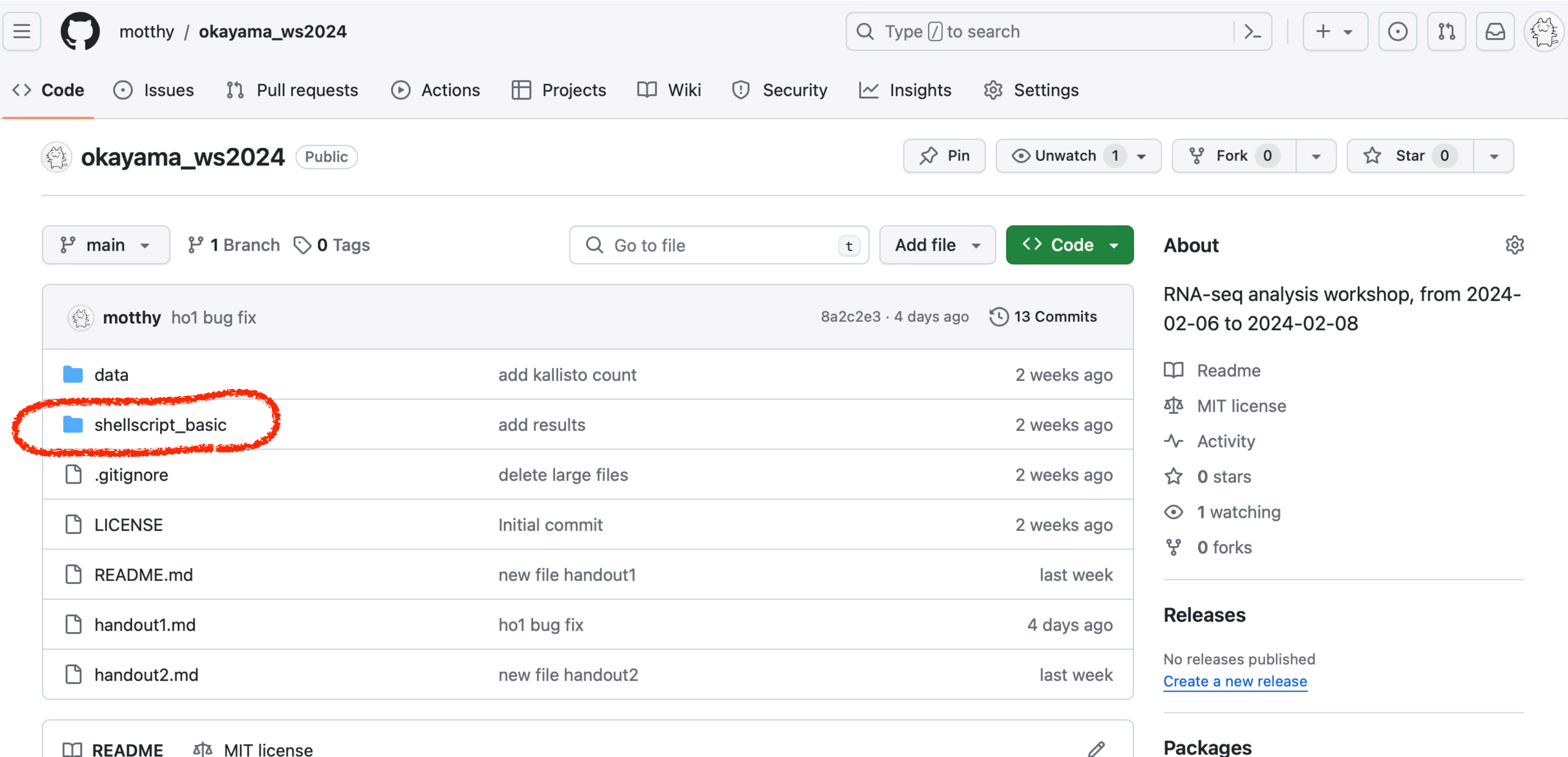
chmodコマンドで変更 例: chmod 775 hoge.txt rwxrwxr-x = 「その他」 書き込み不可

CUIに慣れましょう

簡単なシェルスクリプトの練習

https://github.com/motthy/okayama_ws2024/

クリックで開く



The screenshot shows the GitHub repository page for `motthy / okayama_ws2024`. The repository is public and contains a file named `shellsript_basic`, which is circled in red. The repository description is "RNA-seq analysis workshop, from 2024-02-06 to 2024-02-08".

The repository contains the following files and folders:

File/Folder	Commit Message	Commit Date
data	add kallisto count	2 weeks ago
shellsript_basic	add results	2 weeks ago
.gitignore	delete large files	2 weeks ago
LICENSE	Initial commit	2 weeks ago
README.md	new file handout1	last week
handout1.md	ho1 bug fix	4 days ago
handout2.md	new file handout2	last week

The repository also includes a README, MIT license, and activity feed. The repository has 0 stars, 1 watching, and 0 forks.

- 講習の目的と参考書
- 計算環境について
- CUIに慣れましょう
- **遺伝研スパコンの使い方**
- RNA-seq解析パイプラインの説明

遺伝研スパコンの使い方

はじめに

- ・ 遺伝研スパコンを使うひと向け
- ・ 自分のPCを使うひとは（すみませんが） 少々お休み

ハンドアウトにはlocalにcondaで仮想環境を作って解析する方法も載せています

遺伝研スパコンの使い方

ログイン

1. ゲートウェイノード

```
ssh youraccount@gw.ddbj.nig.ac.jp
```

2. インタラクティブノード

```
qlogin
```


遺伝研スパコンの使い方

コマンドの実行方法

- Grid Engineを使う (qlogin、qsub)
 - ✓ インタラクティブノードでコマンド実行
 - ✓ バッチジョブ (コマンドをまとめてスクリプトにして実行)
 - ✓ アレイジョブ (複数のバッチジョブを同時に実行)

遺伝研スパコンの使い方

バッチジョブとアレイジョブ

バッチジョブ スクリプトにして、処理をまとめて実行できる

```
#$ -S /bin/bash
#$ -cwd
#$ -o logs
#$ -e logs
```

```
FQLIST=("DRR357080" "DRR357081" "DRR357082" "DRR357083" "DRR357084")
```

```
for FQ in ${FQLIST[@]};do
    fasterq-dump $FQ -O fastq
done
```

Grid Engineのコマンド

- S スクリプトを実行するインタプリタの指定 (bashで実行)
- cwd ジョブを現在のディレクトリで実行
- o 標準出力の出力先 (ディレクトリも指定できる)
- e 標準エラー出力の出力先 (ディレクトリも指定できる)
- t アレイジョブの指定 (1,2,3,4,5の5個のジョブ)

アレイジョブ スクリプトを一度に多数実行できる

```
#$ -S /bin/bash
#$ -cwd
#$ -t 1-5:1
#$ -o logs
#$ -e logs
```

← 1から5まで, 1ずつ増える=1,2,3,4,5

```
FQLIST=("DRR357080" "DRR357081" "DRR357082" "DRR357083" "DRR357084")
```

```
fasterq-dump ${FQLIST[SGE_TASK_ID - 1]} -O fastq
```

タスクID: -tで指定するジョブの番号

くわしくは、
NIGスーパーコンピュータ「Grid Engineの概要」をご覧ください

https://sc.ddbj.nig.ac.jp/software/grid_engine/

- 講習の目的と参考書
- 計算環境について
- CUIに慣れましょう
- 遺伝研スパコンの使い方
- **RNA-seq解析パイプラインの説明**

RNA-seq解析

✓ fastqの前処理

- ▶ クオリティチェック
- ▶ クオリティによるフィルタリング/トリミング

FastQC

fastp

Trimmomatic

✓ カウントデータの取得

- ▶ リファレンス配列へのマッピング
- ▶ リードカウント

bowtie2

STAR

HISAT2

kallisto

rsem

featureCounts

✓ 発現変動遺伝子を見つける

✓ データの可視化 (PCA、ヒートマップetc.)

RNA-seq解析パイプラインの例

Rhelixa RNAseq解析パイプライン
(遺伝研スパコンで利用可能)

改訂版RNA-Seqデータ解析 WETラボのための超鉄板レシピ Chapter 3より (羊土社)

前処理

fastq(生データ)

FastQC

Trimmomatic

fastq

fastp

FastQ Screen

fastq

FastQC

Multi QC

fastq

リファレンストランスクリプトーム
データ使用のためQCなし

マッピング

HISAT2

ゲノムにマッピング

HISAT2

STAR

トランスクリプトームに
マッピング

リード
カウント

featureCounts

StringTie

Ballgown

rsem

kallisto

salmon

その他の解析パイプラインツール

ウェブツール・有償ツール

[README](#) [License](#)

DOI [10.5281/zenodo.4718200](https://doi.org/10.5281/zenodo.4718200)

ikra v2.0.1 -RNAseq pipeline centered on Salmon-


A gene expression table (gene x sample) is automatically created from the experiment matrix. The output can be used as an input of [idep](#). Ikra is an RNAseq pipeline centered on [salmon](#).

[日本語ドキュメントはこちら](#)


Note that sra-tools has to be installed locally. This is up to NCBI's tool upgrade. Please install sra-tools (>=2.10.7).

Usage

```
Usage: ikra.sh experiment_table.csv species \
  [--test, --fastq, --help, --without-docker, --udocker --protein-coding] \
  [--threads [VALUE]][--output [VALUE]]\
  [--suffix_PE_1 [VALUE]][--suffix_PE_2 [VALUE]]
args
```



<https://github.com/yyoshiaki/ikra>

 **RaNA-seq**

Home Analysis Pipeline Help Contact

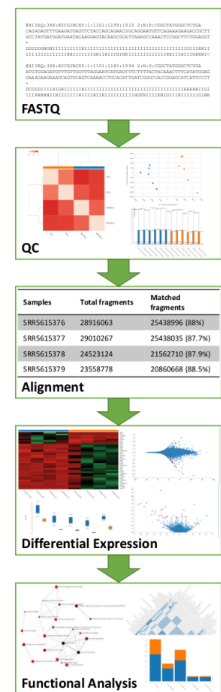
Unregistered user, [login](#) or [register](#)

Home

RaNA-Seq is an open bioinformatics tool for the quick analysis of RNA-Seq data. It performs a full analysis in minutes quantifying FASTQ files, calculating quality control metrics, running differential expression analyses and enabling the interpretation of results with functional analyses. Our analysis pipeline integrates cutting edge bioinformatics tools and simplify its application with a friendly Web interface designed for non-experienced users in these analyses. Each analysis can be customized setting up input parameters and applies generally accepted and reproducible protocols. Analysis results are presented as interactive graphics and reports, ready for their interpretation and publication.

Main features of RaNA-Seq:

- Full RNA-Seq analysis from the FASTQ file to the functional analysis of results
- Stand alone website without any software installation required
- Connected with the ENA repository
- Auto-configuration of input parameters oriented to non-experienced users
- Quality control of input samples with several graphs and metrics
- Customization of analyses with input parameters and several differential expression methods
- Functional over-representation analysis of results
- Gene Set Enrichment Analysis of results (GSEA)
- Generation of reports with full information about the analysis, ready for its inclusion in publications
- Presentation of results with interactive graphics (some examples of [Differential expression results](#), [functional enrichment results](#), [Quality Control metrics](#))
- Generally accepted and reproducible protocols



<https://ranaseq.eu/home>

DRAGEN & BaseSpace Sequence Hub

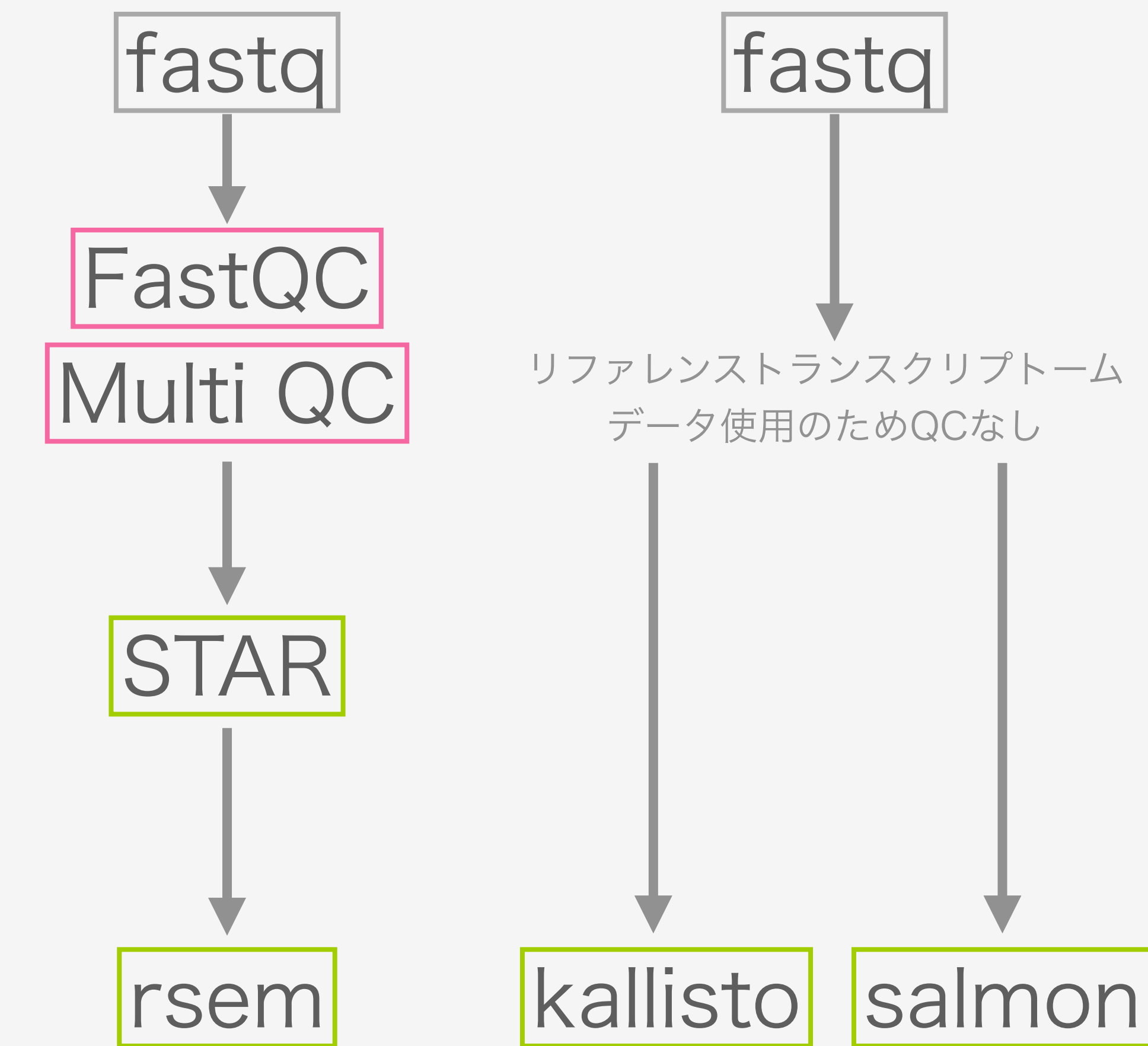
(イルミナ社、有償ツール)

- クラウド利用
- 基本500ポイント/年（追加可能）
- RNA-seq 1サンプルで2ポイント
- パイプライン構築済み
- 爆速

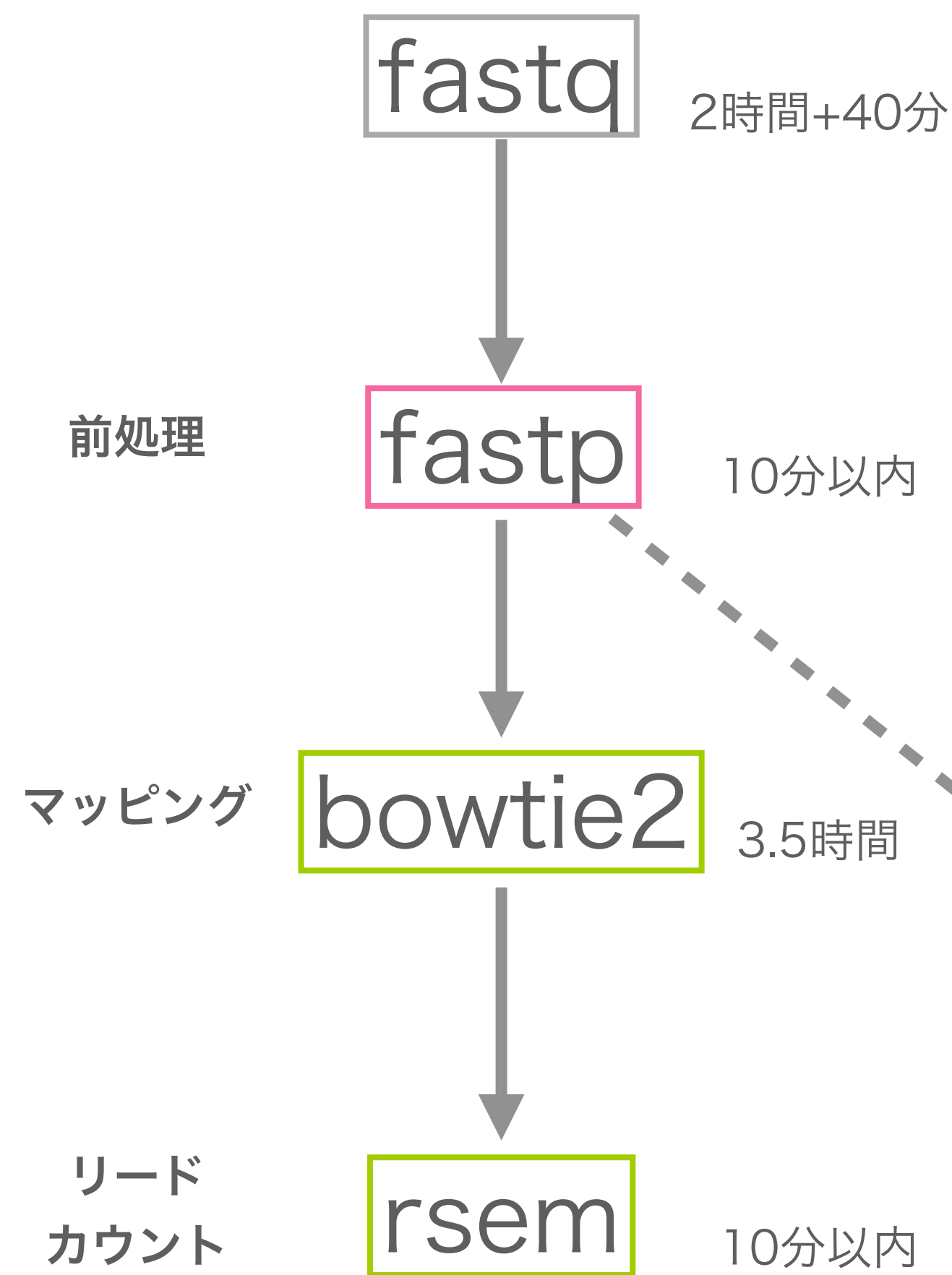
<https://jp.illumina.com/products/by-type/informatics-products/basespace-sequence-hub.html>

この講習で用いるパイプライン

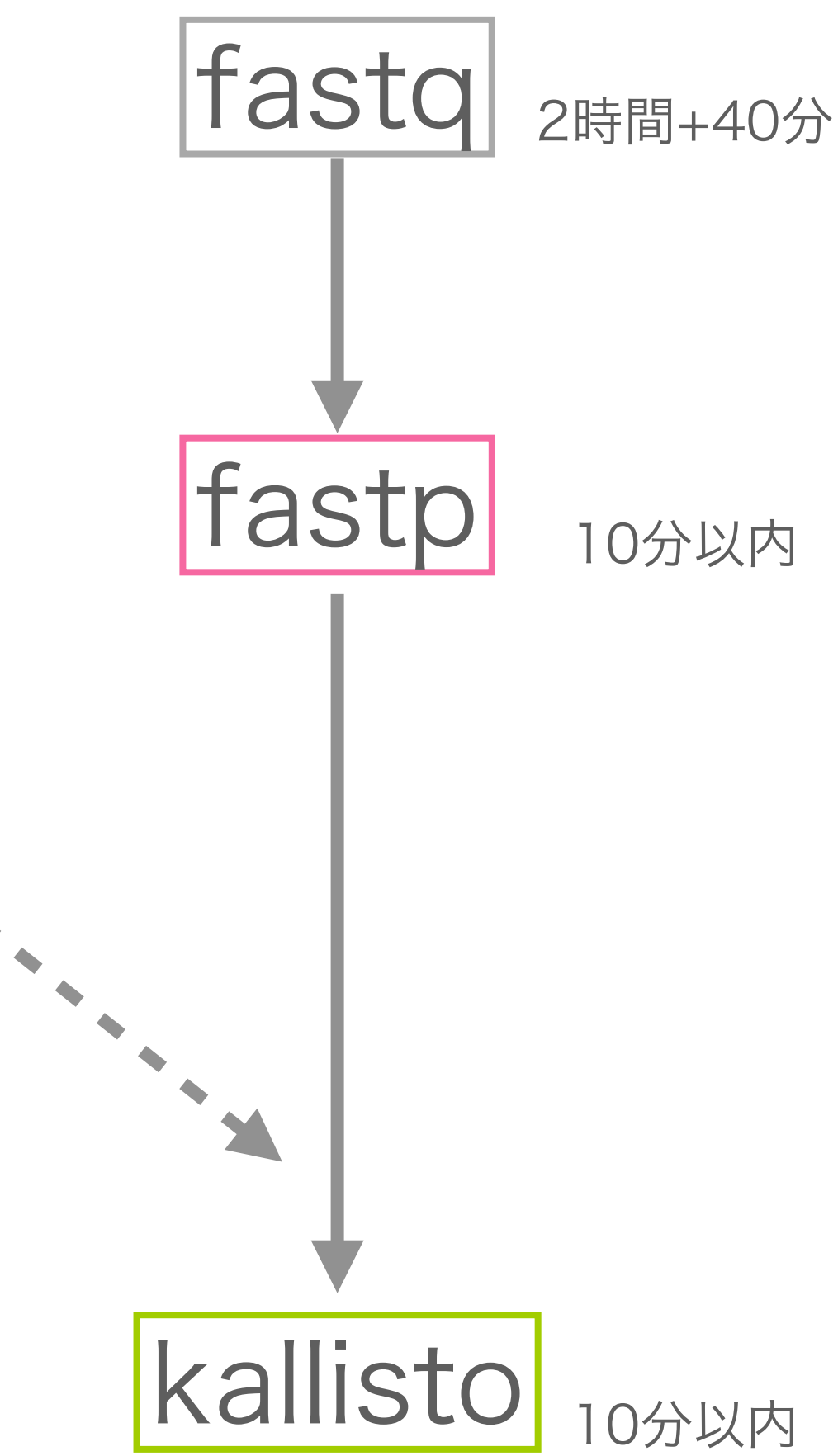
改訂版RNA-Seqデータ解析 WETラボのための超鉄板レシピ Ch.3



リファレンスにマッピングする手法

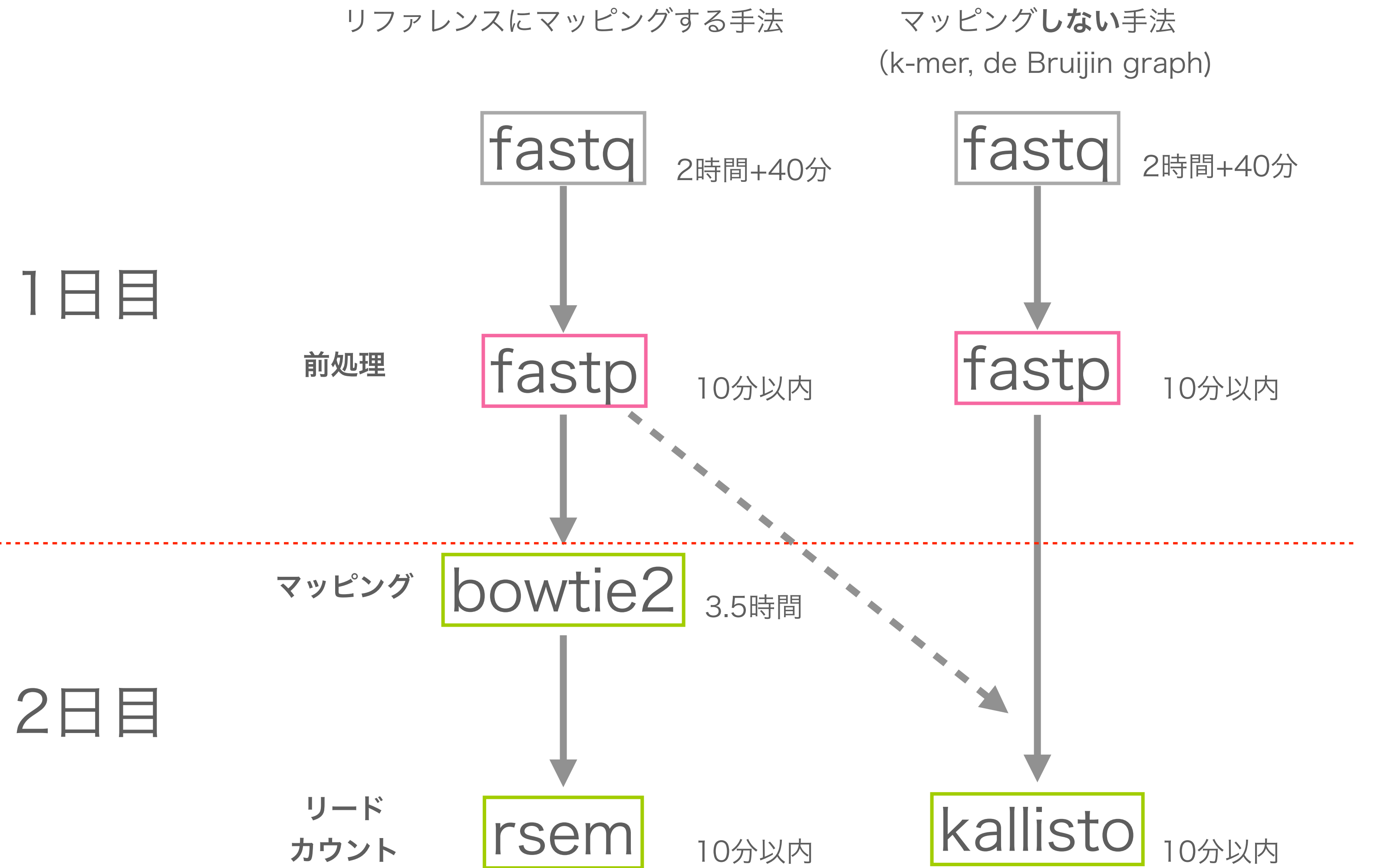


マッピングしない手法
(k-mer, de Bruijn graph)



(スパコンでアレイジョブを実行した場合)

作業予定



では、始めましょう

https://github.com/motthy/okayama_ws2024/

クリックで開く

motthy / okayama_ws2024

Type to search

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

okayama_ws2024 Public

Pin Unwatch 1 Fork 0 Star 0

main 1 Branch 0 Tags

Go to file t Add file Code

About

RNA-seq analysis workshop, from 2024-02-06 to 2024-02-08

Readme MIT license Activity 0 stars 1 watching 0 forks

Releases

No releases published
[Create a new release](#)

Packages

motthy	ho1 bug fix	8a2c2e3 · 4 days ago	13 Commits
data	add kallisto count	2 weeks ago	
shellscript_basic	add results	2 weeks ago	
.gitignore	delete large files	2 weeks ago	
LICENSE	Initial commit	2 weeks ago	
README.md	new file handout1	last week	
handout1.md	ho1 bug fix	4 days ago	
handout2.md	new file handout2	last week	

README MIT license