2019.05.27版

講義資料PDFが講義のページからダウンロード可能です。講義資料の印刷物はありません。課題用のA4一枚はあります。第2回出席予定の持込みPCの方は、当日までにJavaのインストールをしておいてください

機能ゲノム学第1回

[1]大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究プログラム
[2]微生物科学イノベーション連携研究機構
門田幸二（かどた　こうじ）
kadota@iu.a.u-tokyo.ac.jp
http://www.iu.a.u-tokyo.ac.jp/~kadota/

prerequisite

デスクトップ上にhogeフォルダを作成してください。①本科目のウェブページにいき、②今日の講義で用いるファイルをhogeフォルダ上にダウンロードしておいてください。③講義資料PDFをざっと眺めておきましょう

9.機能ゲノム学

①

授業の目標・概要

細胞中で発現している全転写物（トランスクリプトーム）解析手法について、特に塩基配列解析部分を中心に解説します。また、Rのスキルアップを目指します。

担当教員

門田幸二 （東大・農・アグリバイオ / 准教授）

お知らせ

講義では、Rの様々なパッケージを利用します。持ち込み用PC利用希望者は インストール | について を参考にしてR本体および必要なパッケージ群を必ずインストールしておいてください。

フリーソフトウェアRの基本的な利用法を習得済みであることを前提として行いますので、 基本的な利用方法 を参考にして基礎的な事柄を理解しておいてください。

参考図書

門田幸二 著 (金明哲 編)、「シリーズ Useful R ⑦ トランスクリプトーム解析」、共立出版、2014。ISBN:978-4-320-12370-0

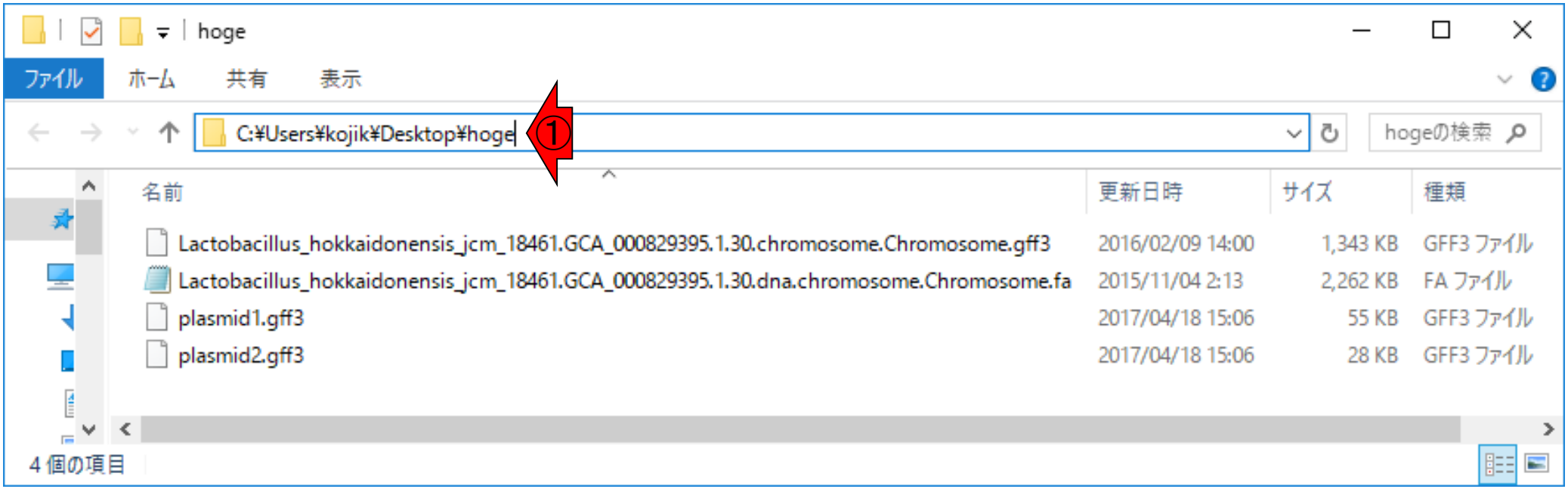
坊農秀雄 著、生命科学データ解析、MEDSi、2017

講義日程（2019年度）

1. 2019年05月27日
   - 講義資料PDF
   - .gff3ファイル (約1.3MB)
   - .faファイル (約2.2MB)
   - (Rで)塩基配列解析
   - (Rで)マイクロアレイデータ解析
   - plasmid1.gff3(課題用)
   - plasmid2.gff3(課題用)
   - de Lannoy et al., F1000Res., 2017
   - Garalde et al., Nat Methods, 2018
   - RNACocktail : Sahraeian et al., Nat Commun., 2017
2. 2019年06月03日
3. 2019年06月10日
4. 2019年06月17日

確認

①デスクトップ上のhogeフォルダ内がこんな感じになっていれば最低限OK



講義予定

細胞中で発現している全転写物（トランスクリプトーム）解析手法について、特に塩基配列解析部分を中心に解説します。また、Rのスキルアップを目指します。

- 第1回（2019年05月27日）
  - 原理、データ解析の概要
  - トランスクリプトーム配列解析、公共データベース（DB）
- 第2回（2019年06月03日）
  - 公共DB関連のTips、FASTQ、ウェブブラウザに注意
  - 前処理（Preprocessing） or Quality Control (QC)
- 第3回（2019年06月10日）
  - クオリティコントロール（QC）、FastQCとFaQCs
  - マッピング（アラインメント）
- 第4回（2019年06月17日）
  - ゲノム配列へのマッピング（アラインメント）の続き
  - カウント情報取得、アノテーション情報の有無、オプションの違い

Contents

- トランスクリプトーム解析技術の原理や特徴
  - RNA-seq（Illuminaの場合）、遺伝子≠転写物
- RNA-seqデータ解析のイメージ
  - マッピング→新規転写物の同定
- 様々な解析目的
  - 転写物配列取得、手法比較論文の紹介、ウェブツール
  - データ解析の全体像（入出力の関係や代表的なツール）
- アノテーションファイルの読み込みと課題1
  - Rで転写物配列取得のイントロ
- Rで転写物配列取得と課題2
  - アノテーションファイルとゲノム情報ファイルから
- 公共データベース
  - NGS全体（NCBI SRA, EMBL-EBI ENA, DDBJ SRA）
  - DRAの概要、クオリティスコアなど

イントロダクション

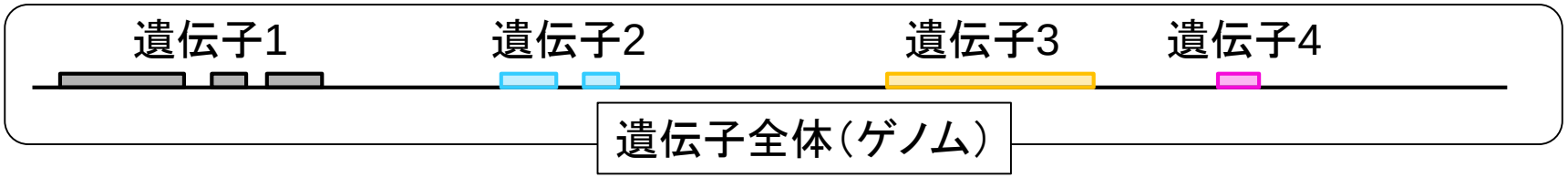
調べたいサンプルでゲノム中のどの領域が、どういう時期に、どの程度転写されている(発現している)かを調べるのがトランスクリプトーム解析。**遺伝子発現解析**や**発現解析**は、トランスクリプトーム解析の一部

- トランスクリプトームとは
  - ある特定の状態の組織や細胞中に存在する全RNA(転写物、transcripts)の総体
- 様々なトランスクリプトーム解析技術
  - **マイクロアレイ**(配列既知の生物種)
    - Affymetrix GeneChip、Illumina BeadArrayなど
  - 配列決定に基づく方法(配列未知でもよい)
    - EST、SAGE、CAGE、**RNA-seq**など

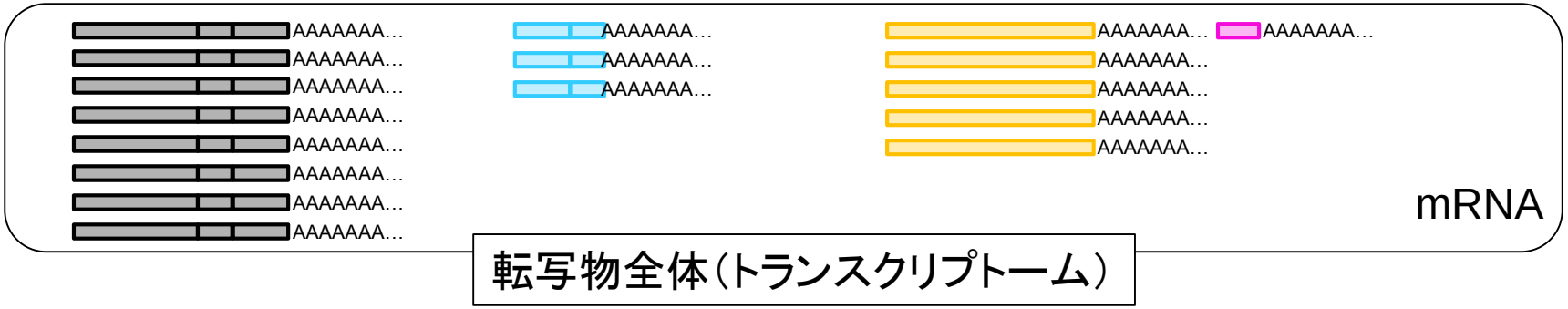
働いているRNAの種類や量を調べるのが目的

トランスクリプトーム解析

- ある状態のあるサンプル（例：目）のあるゲノムの領域



・どの染色体上のどの領域にどの遺伝子があるかは調べる個体（例：ヒト）が同じなら不変（目だろうが心臓だろうが…）

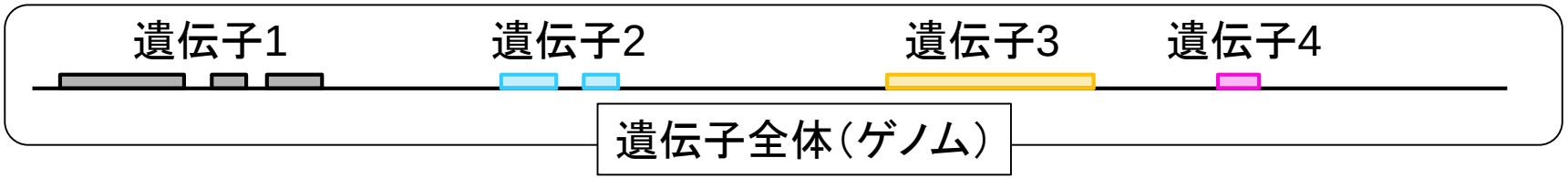


・遺伝子1は沢山転写されている（発現している）
・遺伝子4はごくわずかしか転写されてない
・…

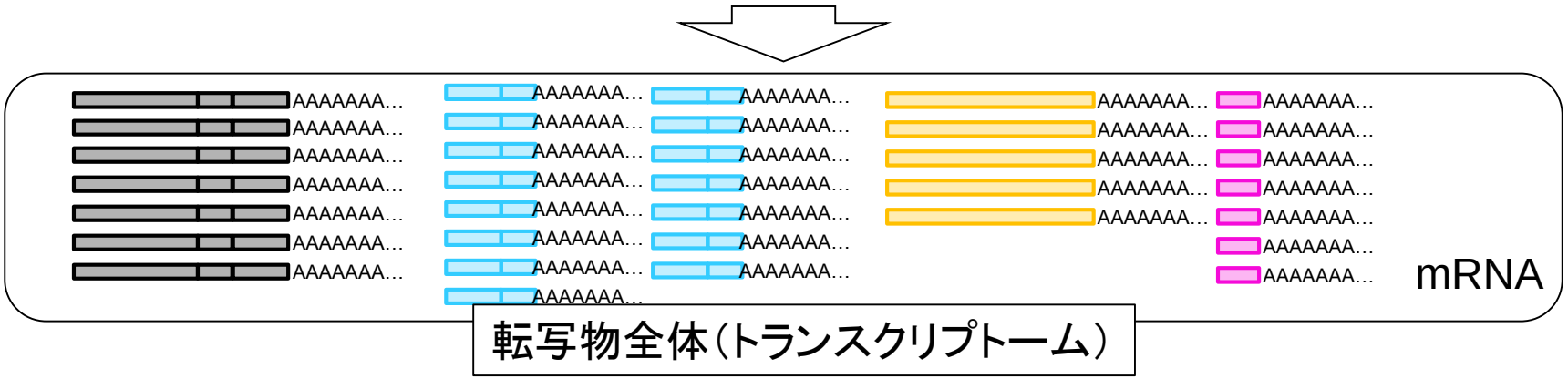
トランスクリプトーム解析

働いているRNAの種類や量を調べるのが目的

- ある状態のあるサンプル（例：目）のあるゲノムの領域



・どの染色体上のどの領域にどの遺伝子があるかは調べる個体（例：ヒト）が同じなら不変（目だろうが心臓だろうが…）

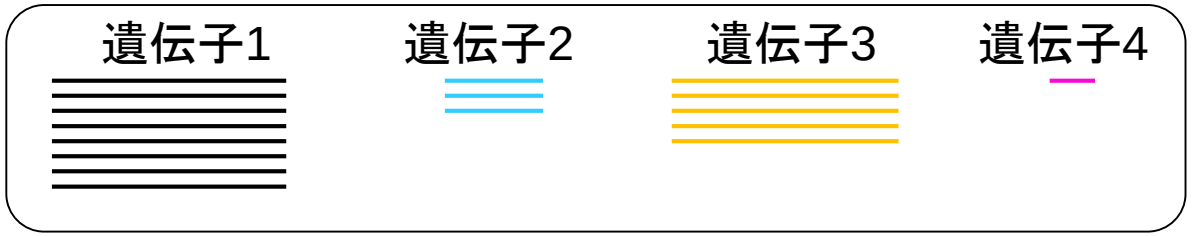


・光刺激に応答して発現亢進するのは遺伝子2と4

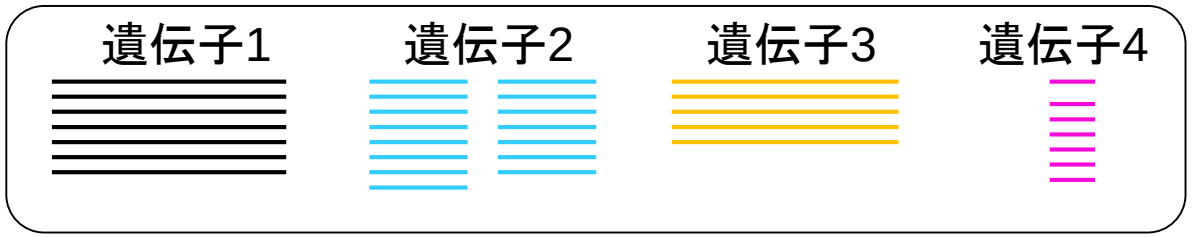
トランスクリプトーム解析

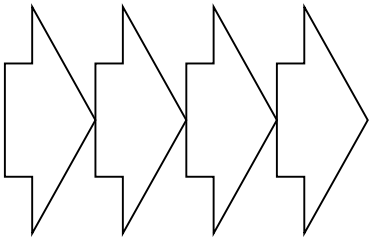
状態の異なる複数サンプルのデータを取得して解析するのが一般的。サンプル間比較。

- 光刺激前（T1）の目のトランスクリプトーム



- 光刺激後（T2）の目のトランスクリプトーム



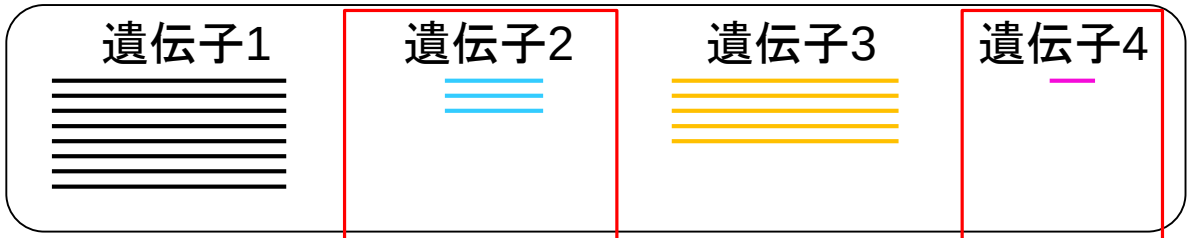


	T1	T2
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	7
…	…	…

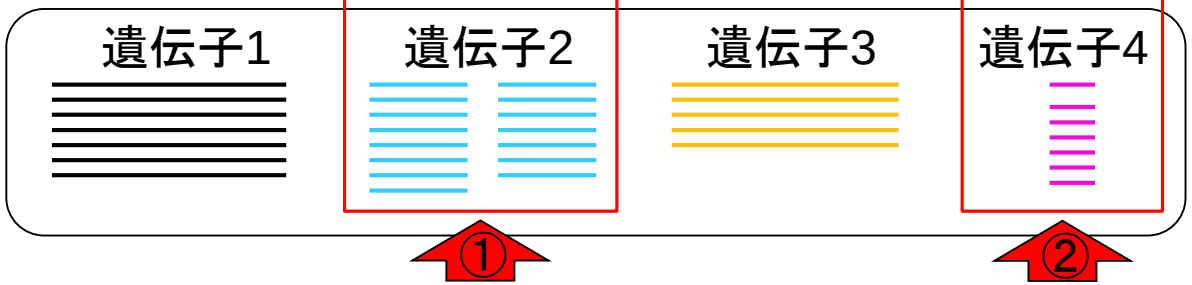
トランスクリプトーム解析

具体的な目的は、①や②の発現変動遺伝子同定など。

- 光刺激前（T1）の目のトランスクリプトーム



- 光刺激後（T2）の目のトランスクリプトーム



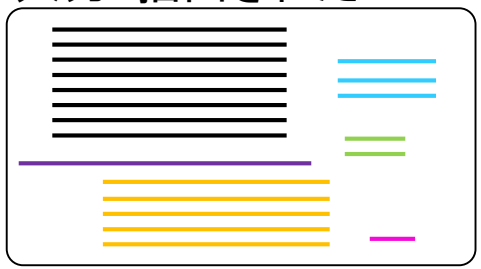
これがいわゆる「遺伝子発現行列」

	T1	T2
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	7
…	…	…

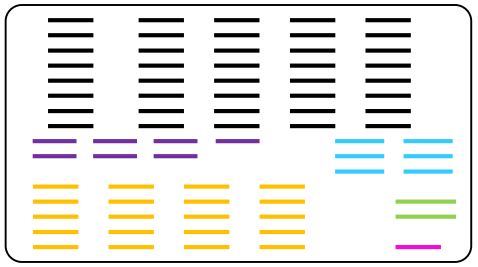
入力：サンプルのRNA
出力：大量塩基配列データ

RNA-seq

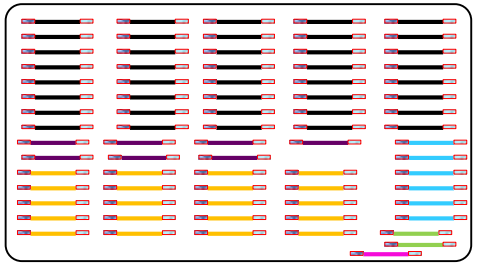
入力：抽出されたRNA



断片化

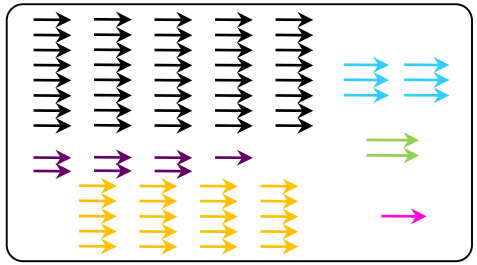


アダプター付加



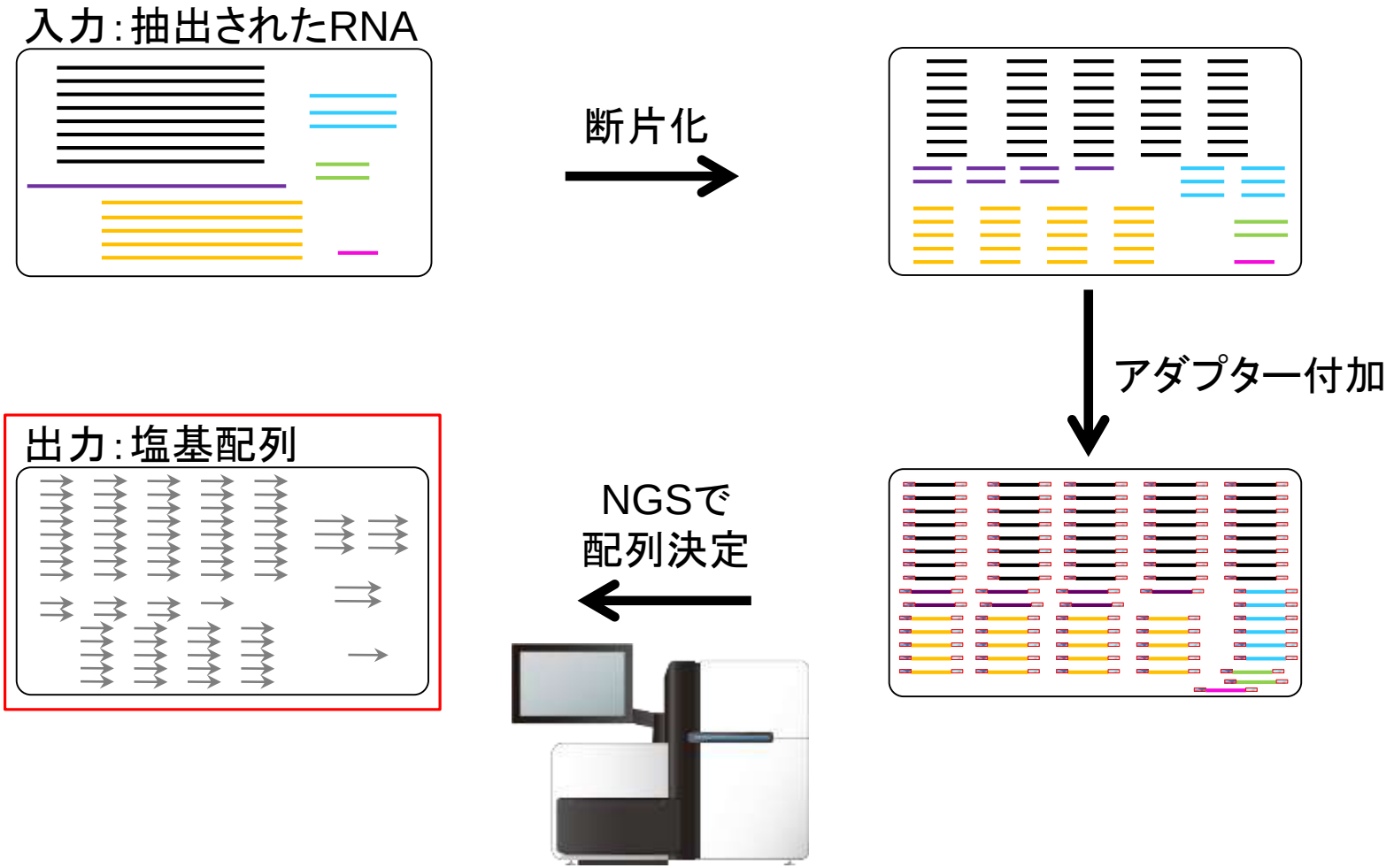
NGSで
配列決定

出力：塩基配列



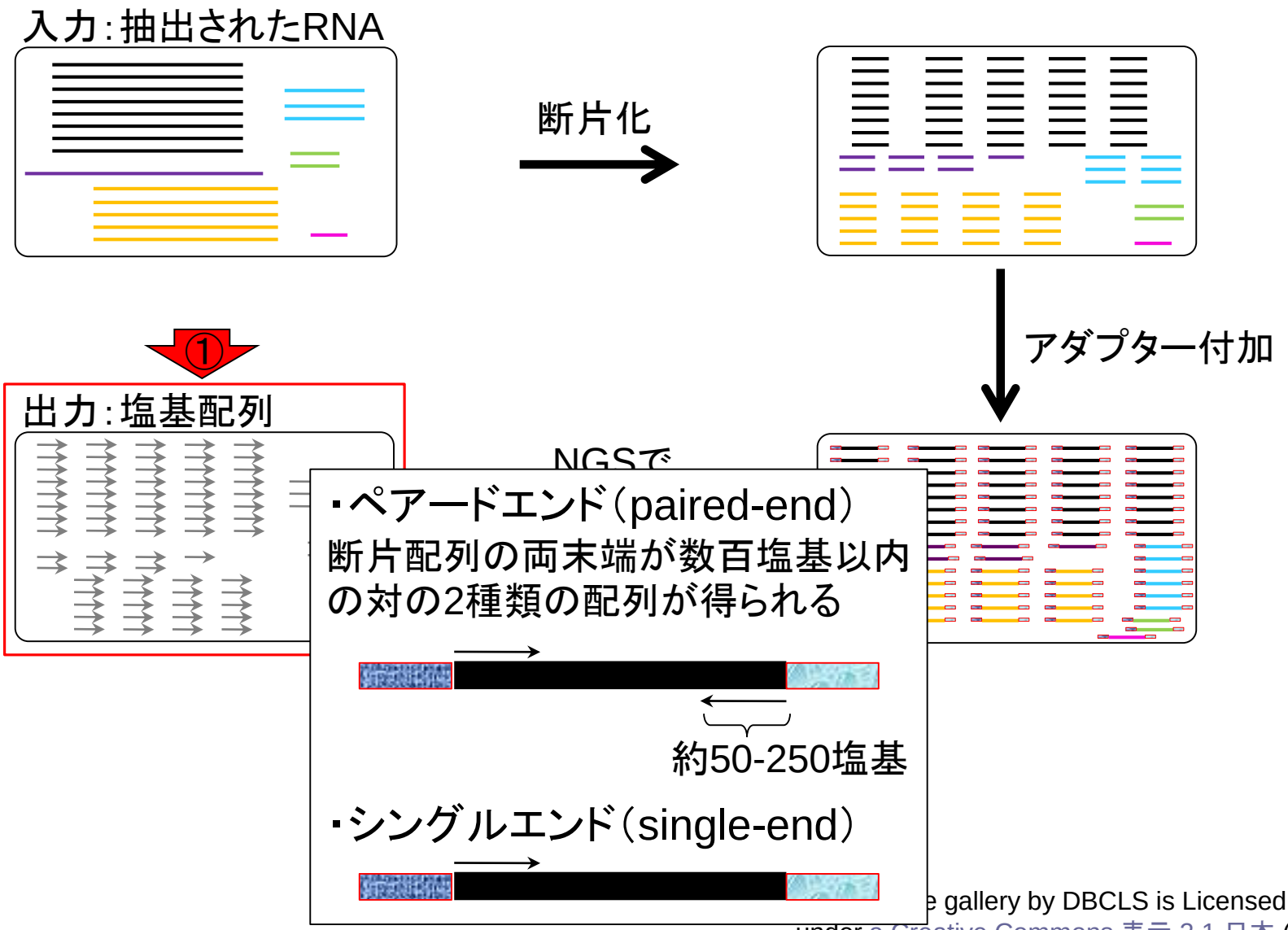
RNA-seq

NGSの出力は、リードと呼ばれる数百塩基程度の配列が延々と続く巨大なファイル。各矢印が1つのリードに相当。この段階では、まだどのリードがどの転写物由来かは不明（なので灰色一色）



RNA-seq

Illuminaの場合は、両側から読むpaired-endと片側のみ読むsingle-endの2つのやり方が存在する。①の出力イメージはsingle-endの場合



Contents

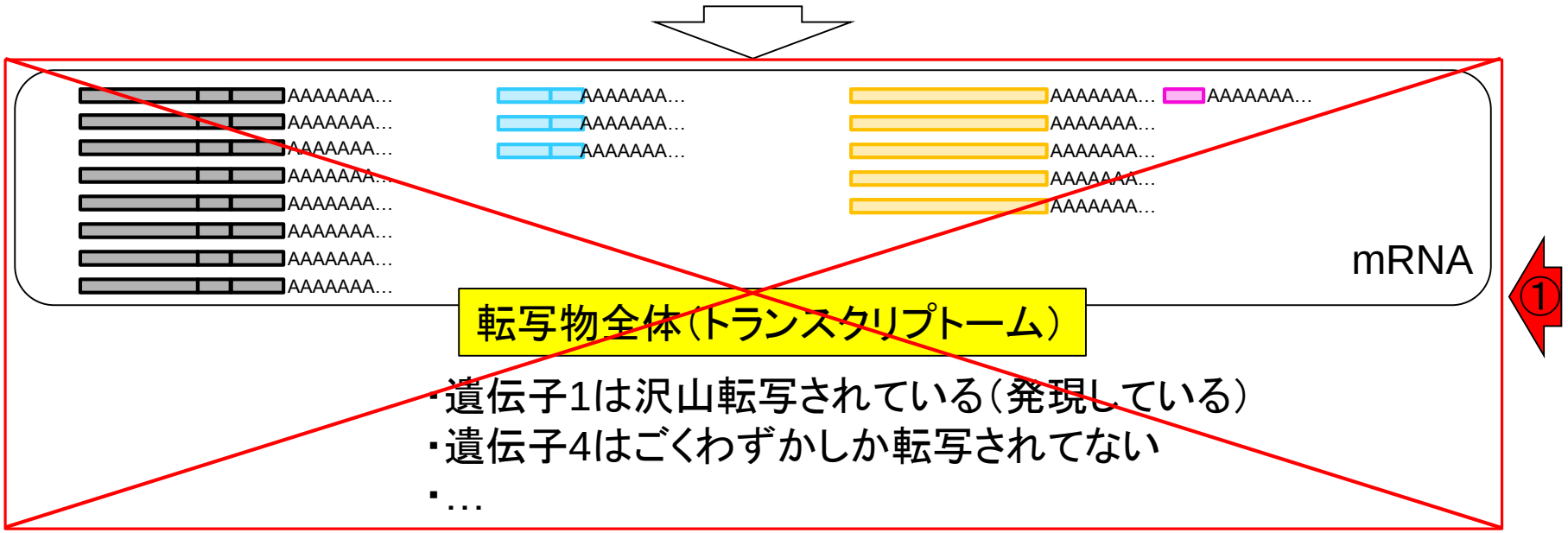
- トランスクリプトーム解析技術の原理や特徴
  - RNA-seq（Illuminaの場合）、遺伝子≠転写物
- RNA-seqデータ解析のイメージ
  - マッピング→新規転写物の同定
- 様々な解析目的
  - 転写物配列取得、手法比較論文の紹介、ウェブツール
  - データ解析の全体像（入出力の関係や代表的なツール）
- アノテーションファイルの読み込みと課題1
  - Rで転写物配列取得のイントロ
- Rで転写物配列取得と課題2
  - アノテーションファイルとゲノム情報ファイルから
- 公共データベース
  - NGS全体（NCBI SRA, EMBL-EBI ENA, DDBJ SRA）
  - DRAの概要、クオリティスコアなど

遺伝子 ≠ 転写物

①赤枠部分の表現は、本当は不正確。昔は実験機器の解像度が事実上**遺伝子レベル**だった。**遺伝子発現解析**という表現はその名残り

- ある状態のあるサンプル（例：目）のあるゲノムの領域

・どの染色体上のどの領域にどの遺伝子があるかは調べる個体（例：ヒト）が同じなら不変（目だろうが心臓だろうが…）



・遺伝子1は沢山転写されている（発現している）
・遺伝子4はごくわずかしか転写されてない
・…

遺伝子 ≠ 転写物

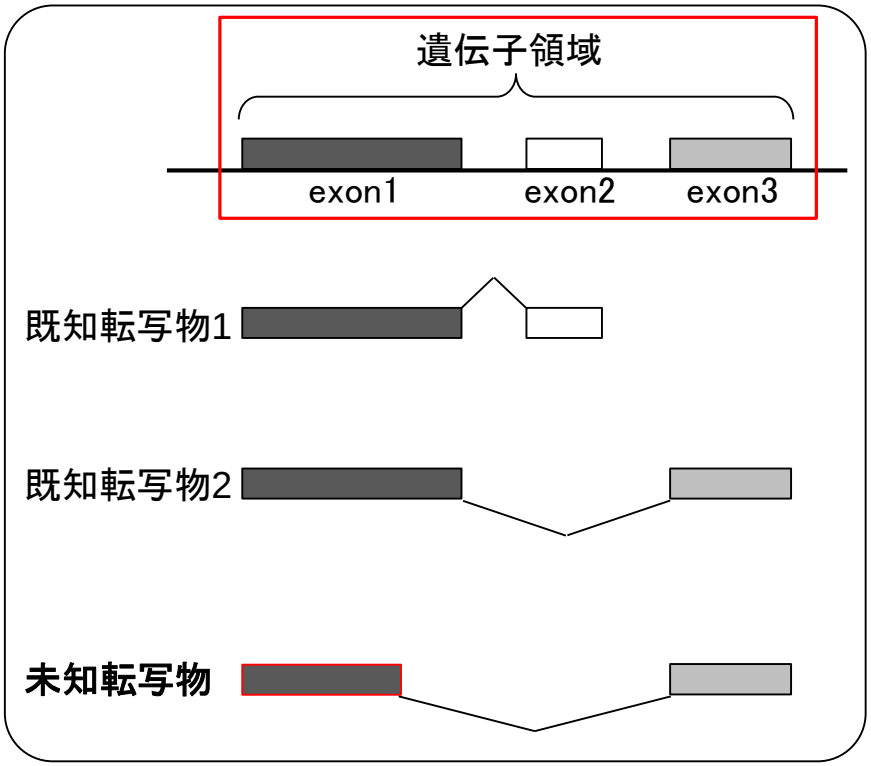
ある遺伝子領域から転写(transcription)されている転写物(transcript)は、1種類とは限らない

- ある状態のあるサンプル(例:目)のあるゲノムの領域

例えば、①遺伝子1の領域では、3種類の真の転写物が存在し、そのうち2種類は既知とする

遺伝子 ≠ 転写物

- ある状態のあるサンプル（例：目）のあるゲノムの領域

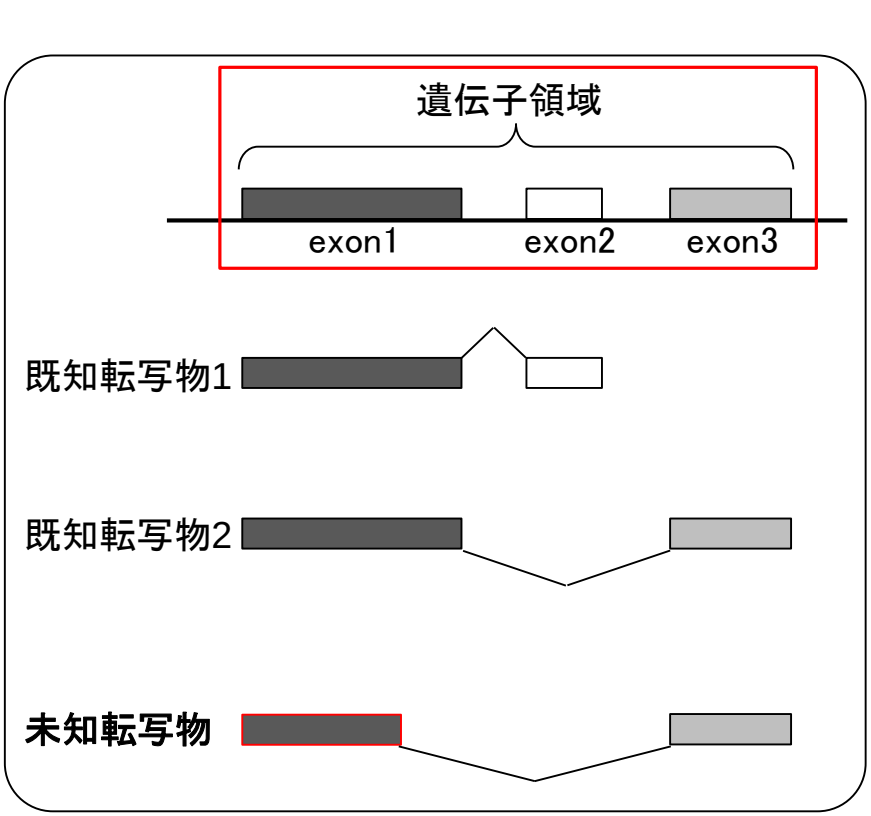


真の転写物情報

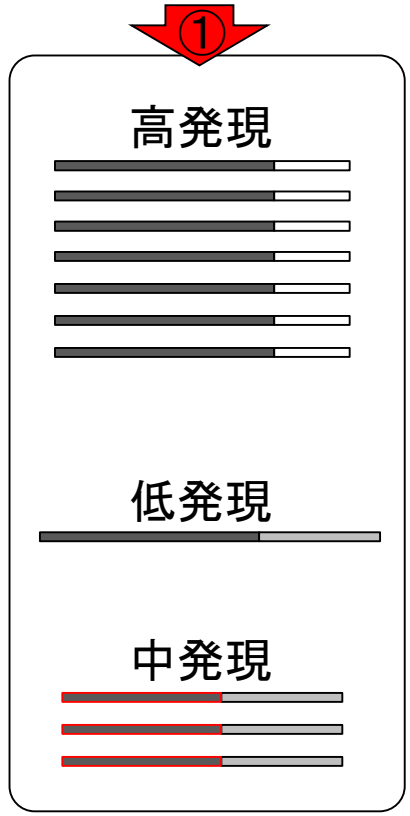
実際の細胞内(例:目のサンプル)での発現情報(働いている度合い)が①のような感じだったとする

遺伝子 ≠ 転写物

- ある状態のあるサンプル(例:目)のあるゲノムの領域



真の転写物情報

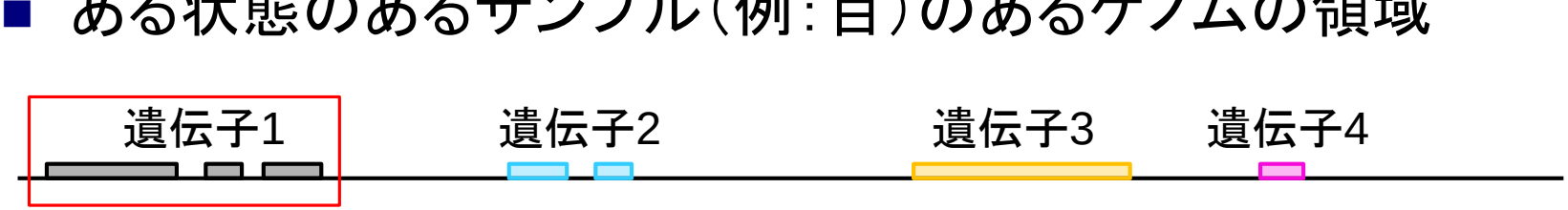


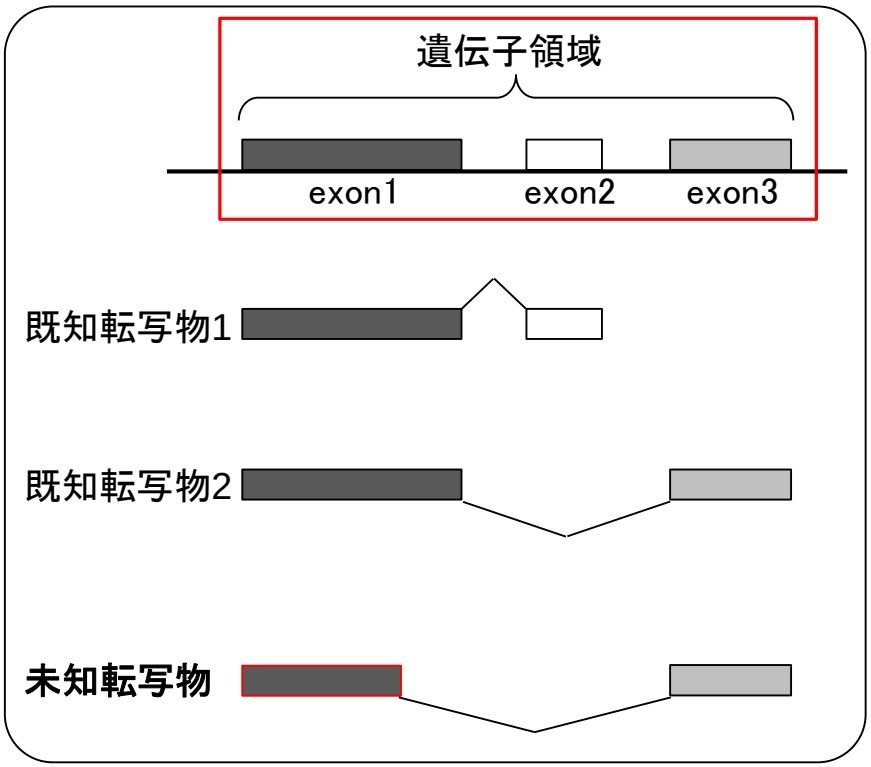
真の発現情報

①NGS機器を用いて転写されているmRNA配列決定（RNA-seq）をした結果のイメージ

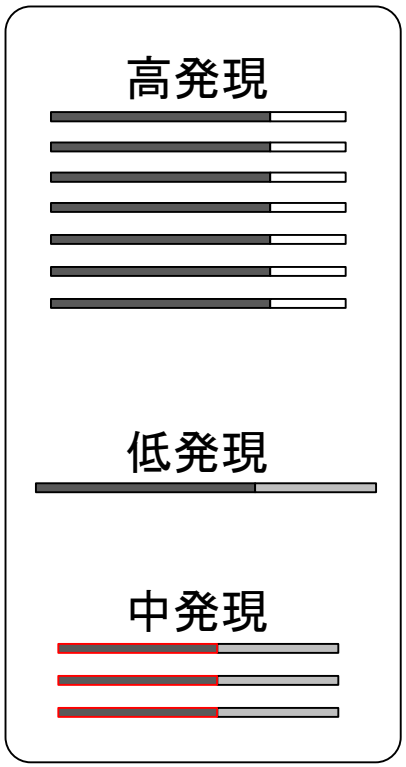
遺伝子 ≠ 転写物

- ある状態のあるサンプル（例：目）のあるゲノムの領域

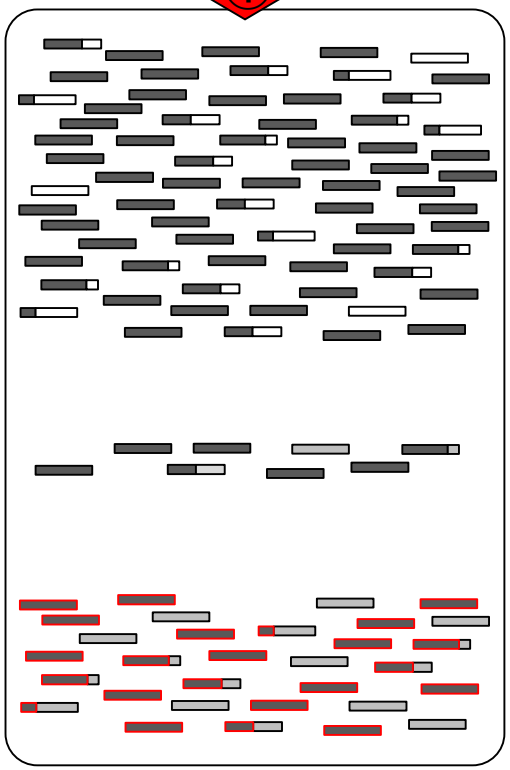




真の転写物情報



真の発現情報



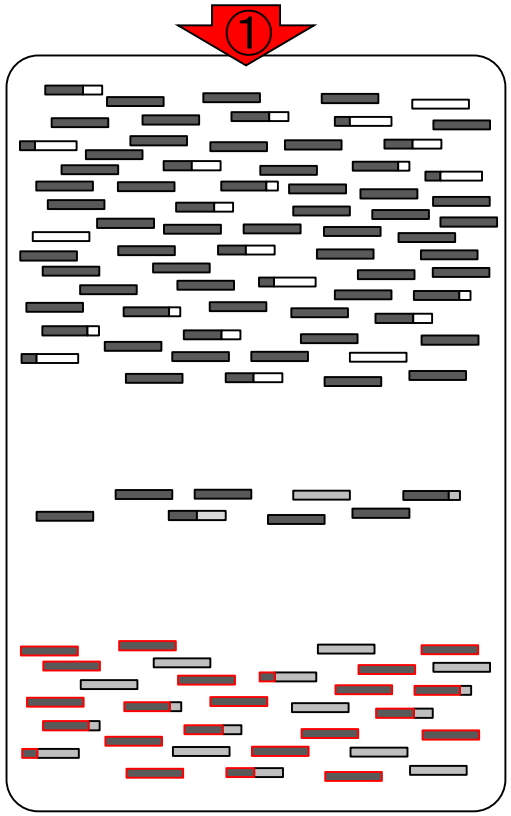
RNA-seqで得られるリード情報（色は不明）

Contents

- トランスクリプトーム解析技術の原理や特徴
  - RNA-seq（Illuminaの場合）、遺伝子≠転写物
- RNA-seqデータ解析のイメージ
  - マッピング→新規転写物の同定
- 様々な解析目的
  - 転写物配列取得、手法比較論文の紹介、ウェブツール
  - データ解析の全体像（入出力の関係や代表的なツール）
- アノテーションファイルの読み込みと課題1
  - Rで転写物配列取得のイントロ
- Rで転写物配列取得と課題2
  - アノテーションファイルとゲノム情報ファイルから
- 公共データベース
  - NGS全体（NCBI SRA, EMBL-EBI ENA, DDBJ SRA）
  - DRAの概要、クオリティスコアなど

データ解析の出発点

トランスクリプトーム（RNA-seq）データ解析の出発点は、①RNA-seqデータファイル、

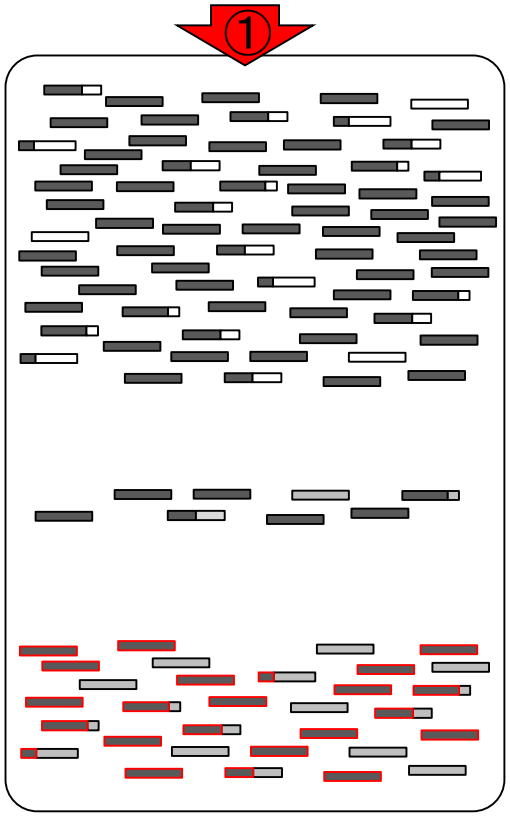


RNA-seqデータ

データ解析の出発点

トランスクリプトーム（RNA-seq）データ解析の出発点は、①RNA-seqデータファイル、②ゲノム配列情報。本当はゲノム配列でなくてもよく、リファレンス配列のほうが正確

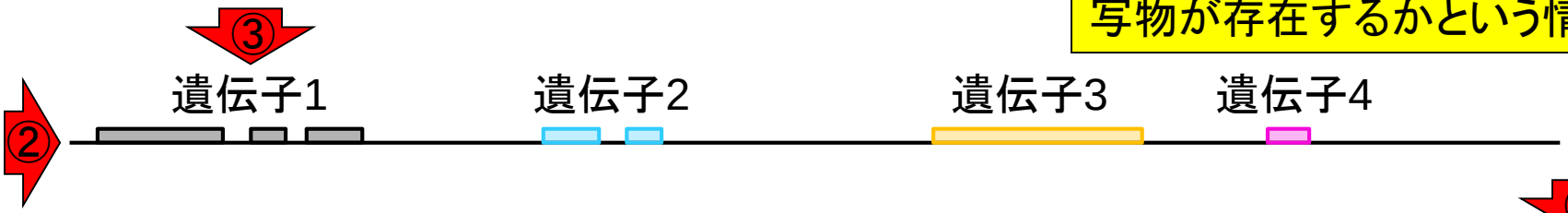
②

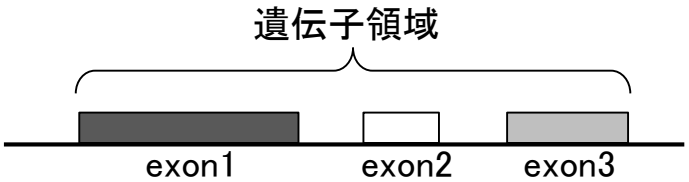


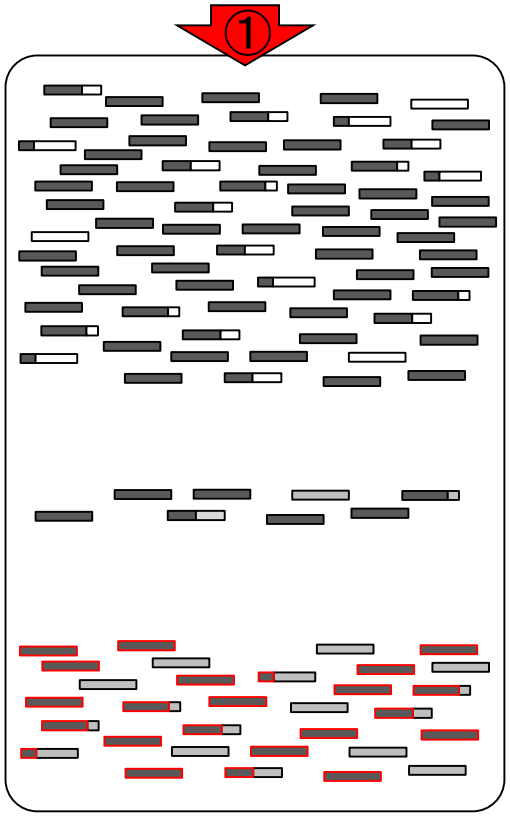
RNA-seqデータ

データ解析の出発点

トランスクリプトーム（RNA-seq）データ解析の出発点は、①RNA-seqデータファイル、②ゲノム配列情報、③アノテーション情報（ゲノム上のどこにどんな遺伝子、exon、転写物が存在するかという情報）



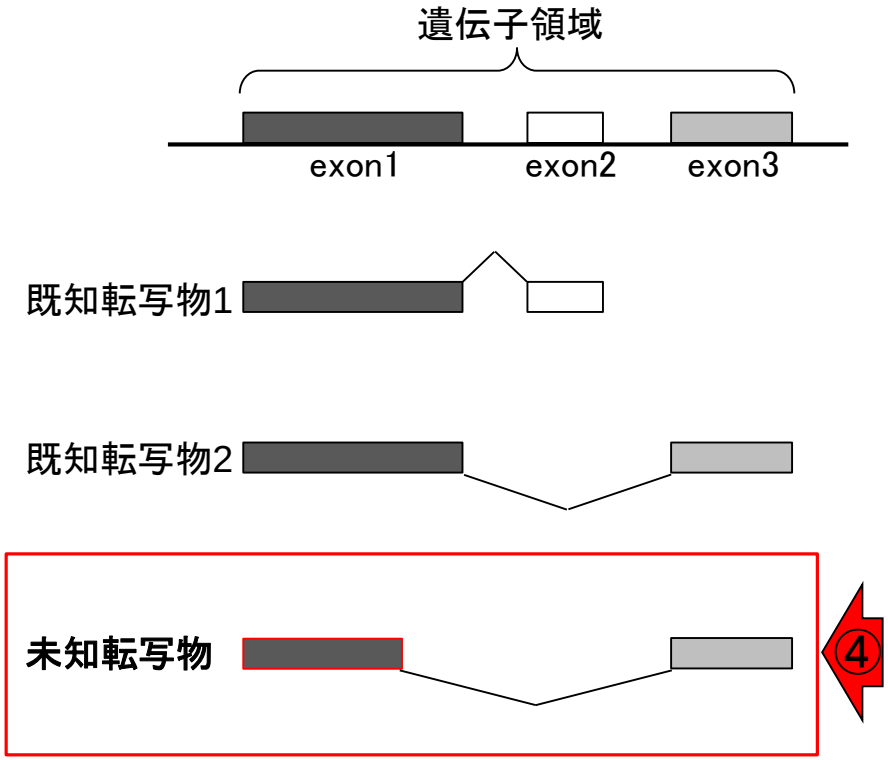


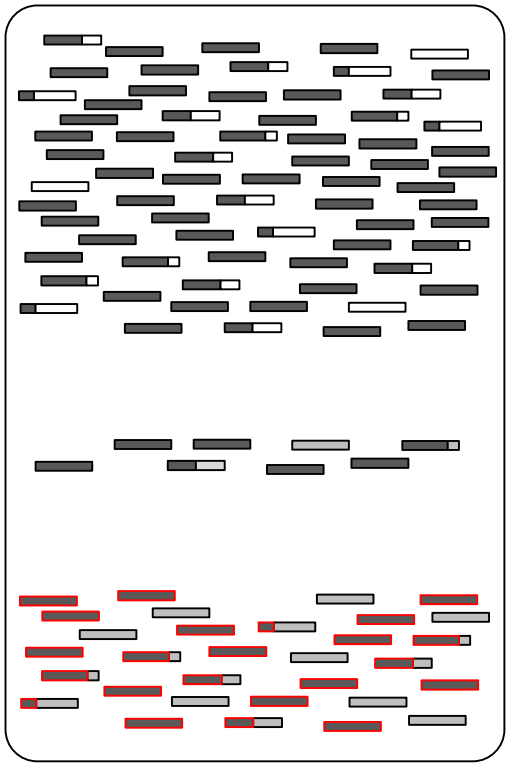


RNA-seqデータ

解析結果のイメージ

①RNA-seqデータ、②ゲノム配列情報、③アノテーション情報を利用して、**④未知転写物（新規isoform）の同定**ができる。



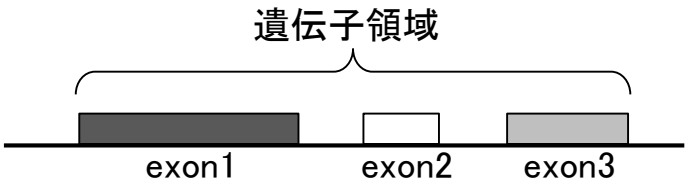


RNA-seqデータ

解析結果のイメージ

①RNA-seqデータ、②ゲノム配列情報、③アノテーション情報を利用して、④未知転写物（新規isoform）の同定ができる。**⑤転写物の発現量（働いている度合い）推定も原理的に可能**

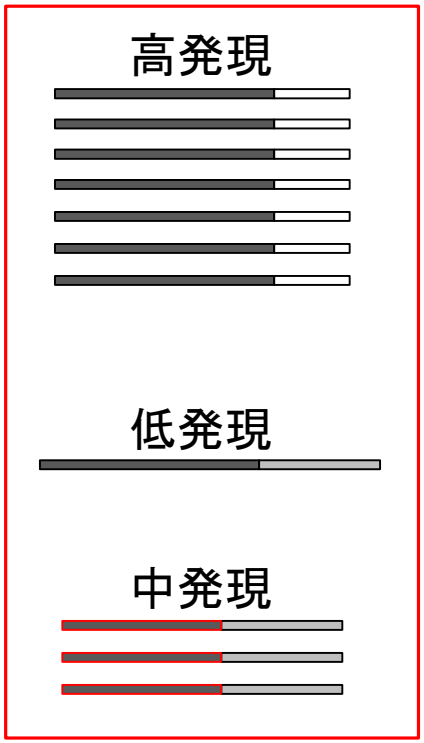
⑤

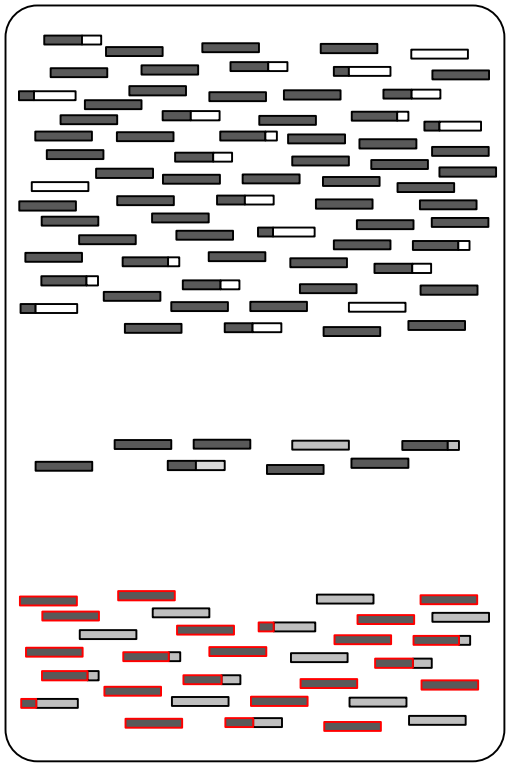


既知転写物1

既知転写物2

未知転写物

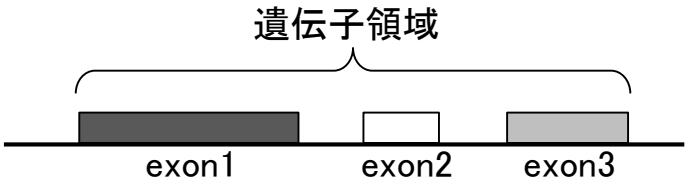


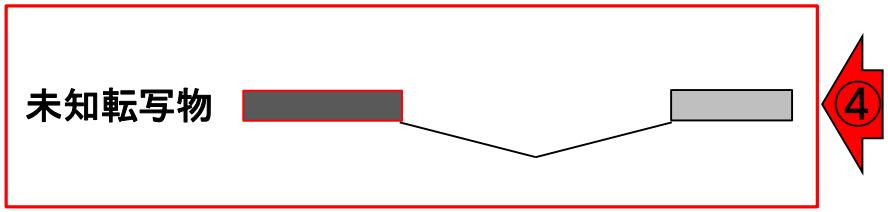


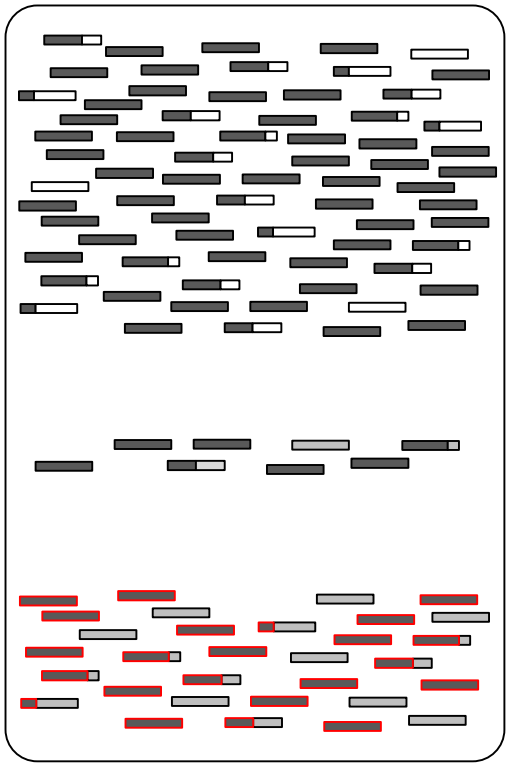
RNA-seqデータ

具体的な戦略は?

①RNA-seqデータ、②ゲノム配列情報、③アノテーション情報を利用して、**④未知転写物（新規isoform）の同定**ができる。



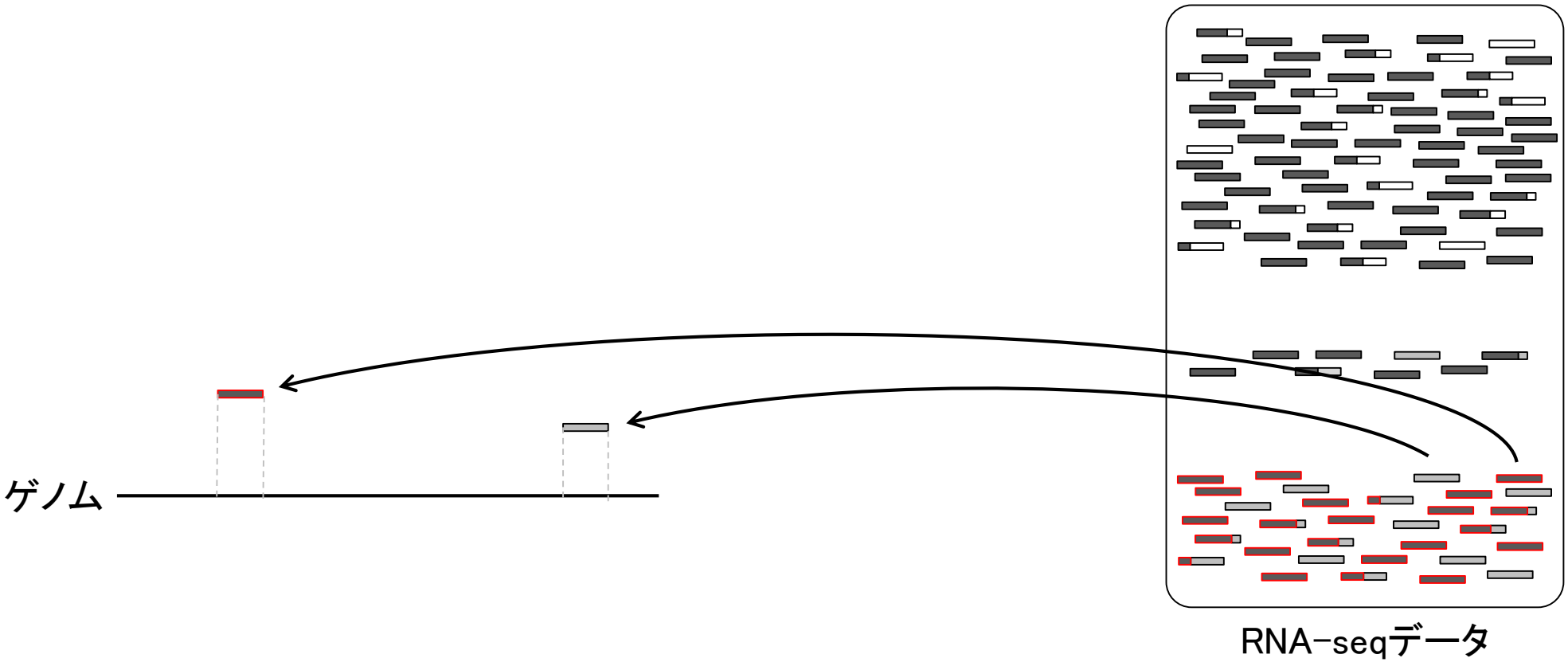




RNA-seqデータ

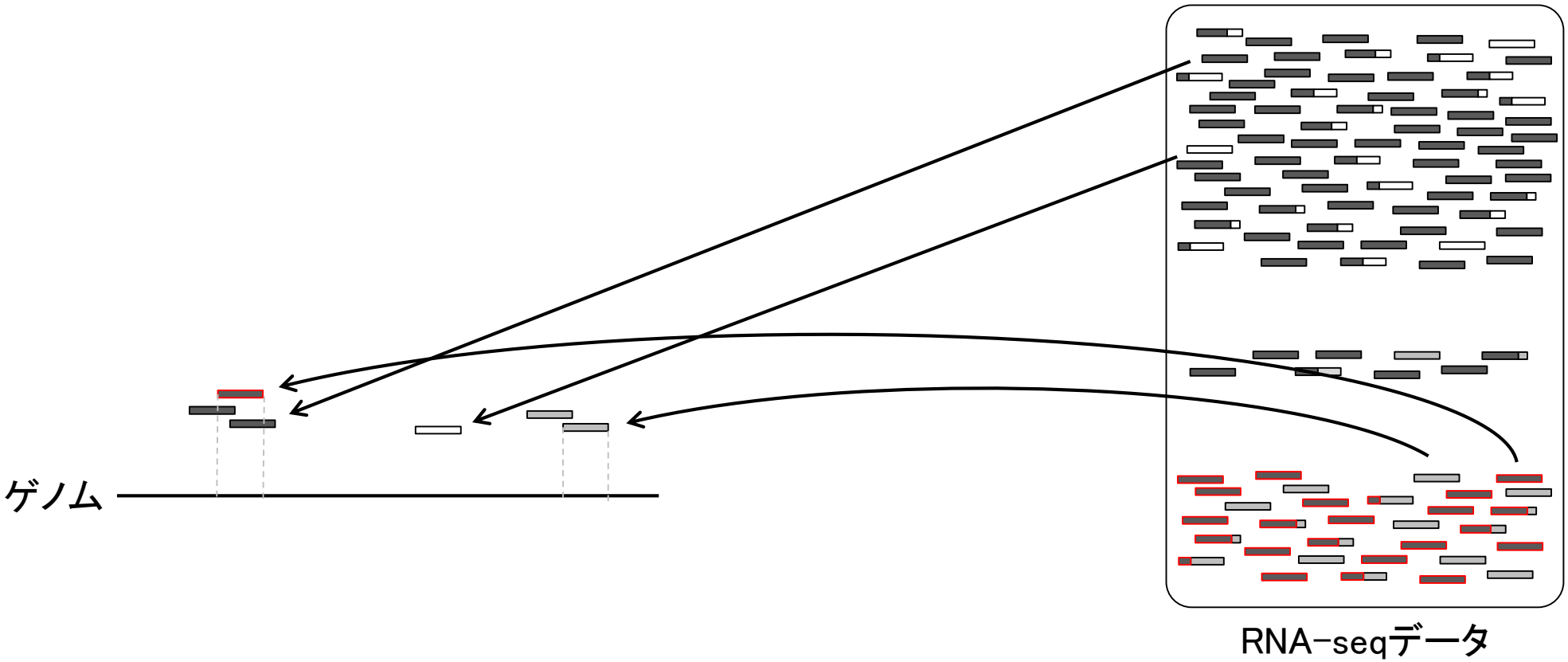
具体的な戦略

RNA-seqデータ中の1本1本のリード（横棒）がゲノム上のどの領域から転写されたのかを調べる。文字列検索と本質的に同じであり、これがマッピングという作業に相当する



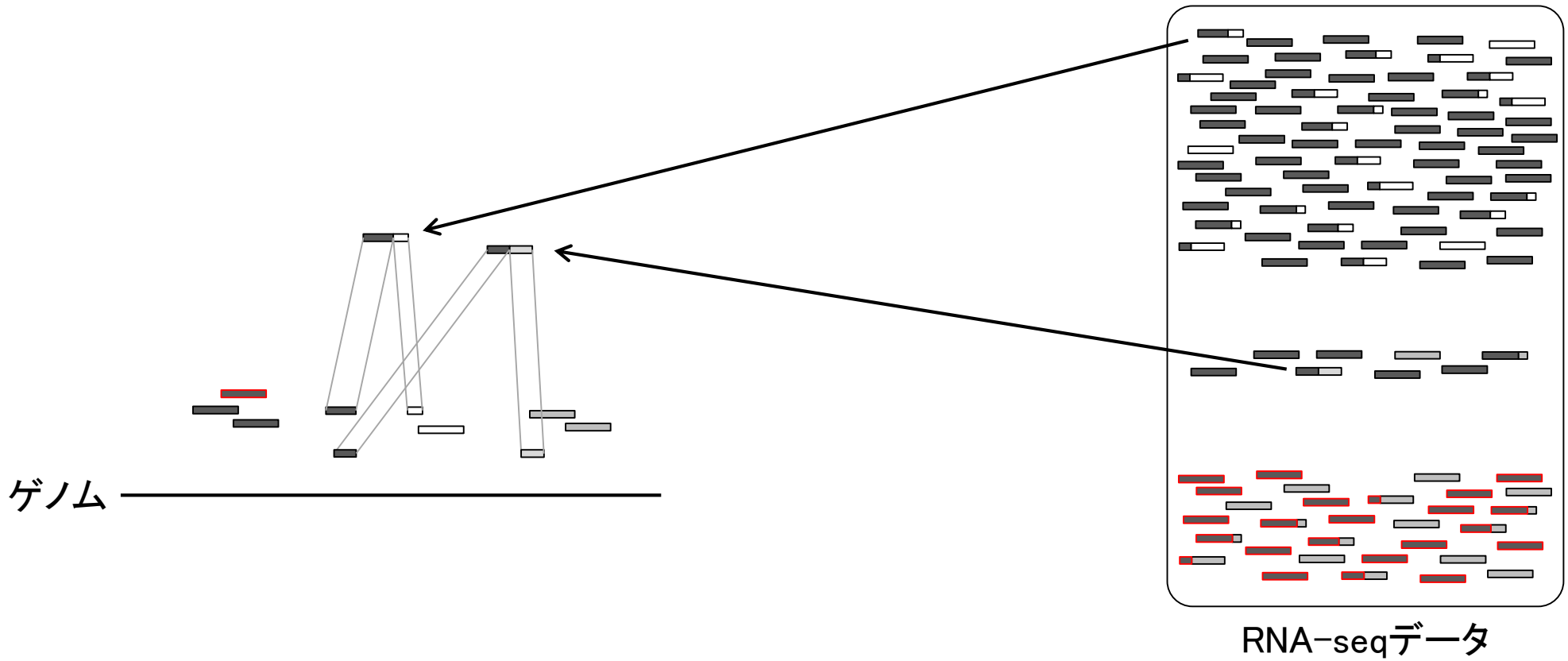
具体的な戦略

RNA-seqデータ中の1本1本のリード（横棒）がゲノム上のどの領域から転写されたのかを調べる。文字列検索と本質的に同じであり、これがマッピングという作業に相当する



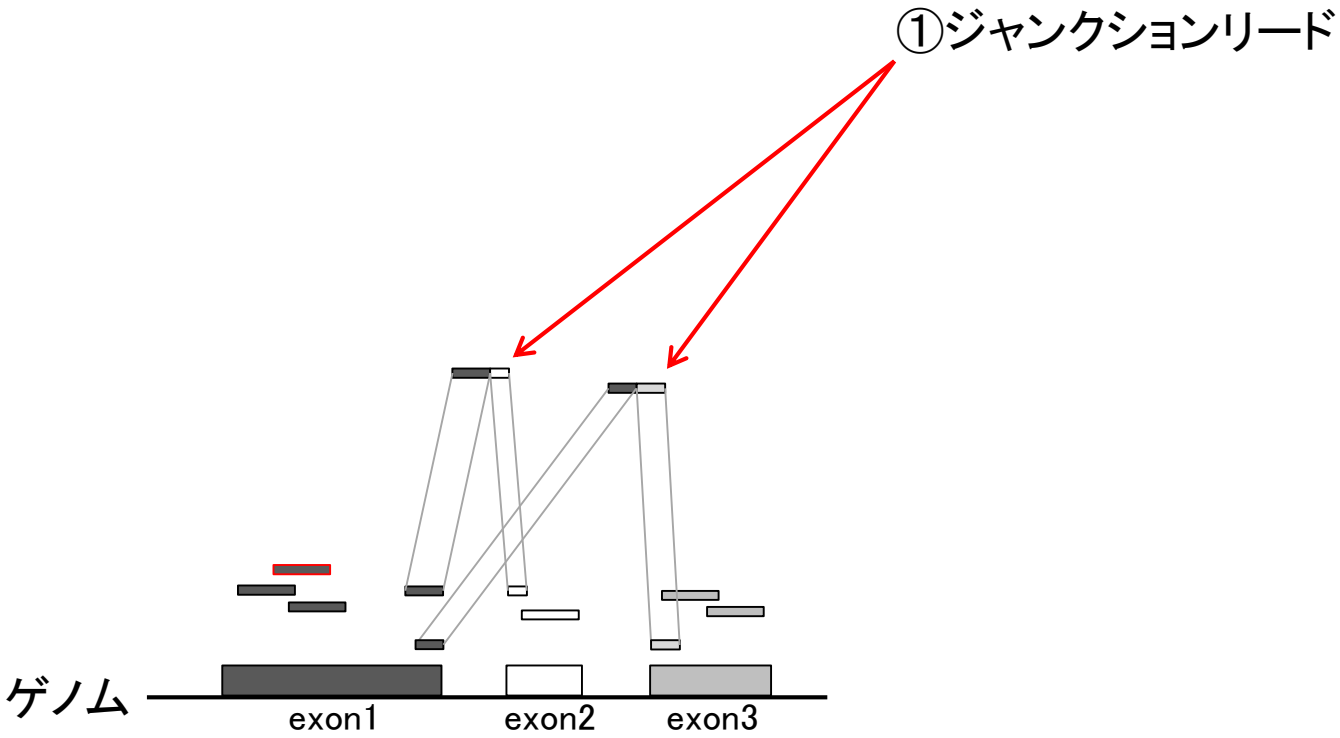
具体的な戦略

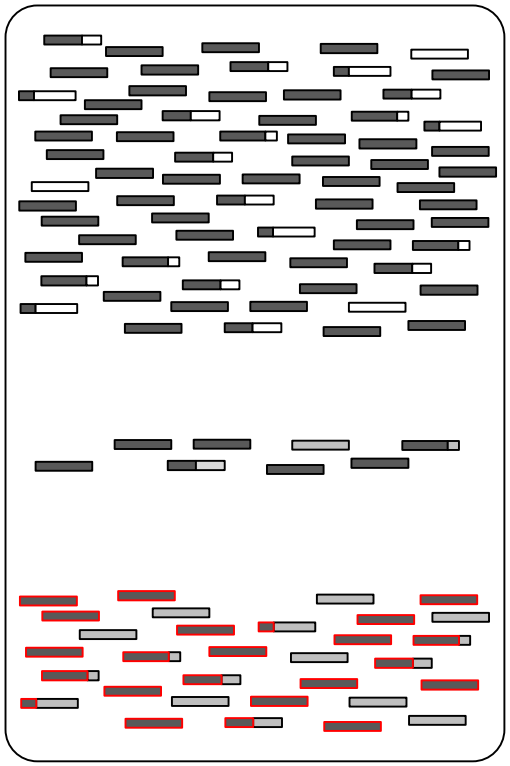
リードの長さが初期は35塩基程度だったが、現在は数百塩基程度まで伸びている。そのおかげで、リードを分割してマップすることもできる



具体的な戦略

分割してマップされたリードは、大抵の場合複数のエクソン（exon）をまたぐリードであり、①ジャンクションリード（junction read）と呼ばれる

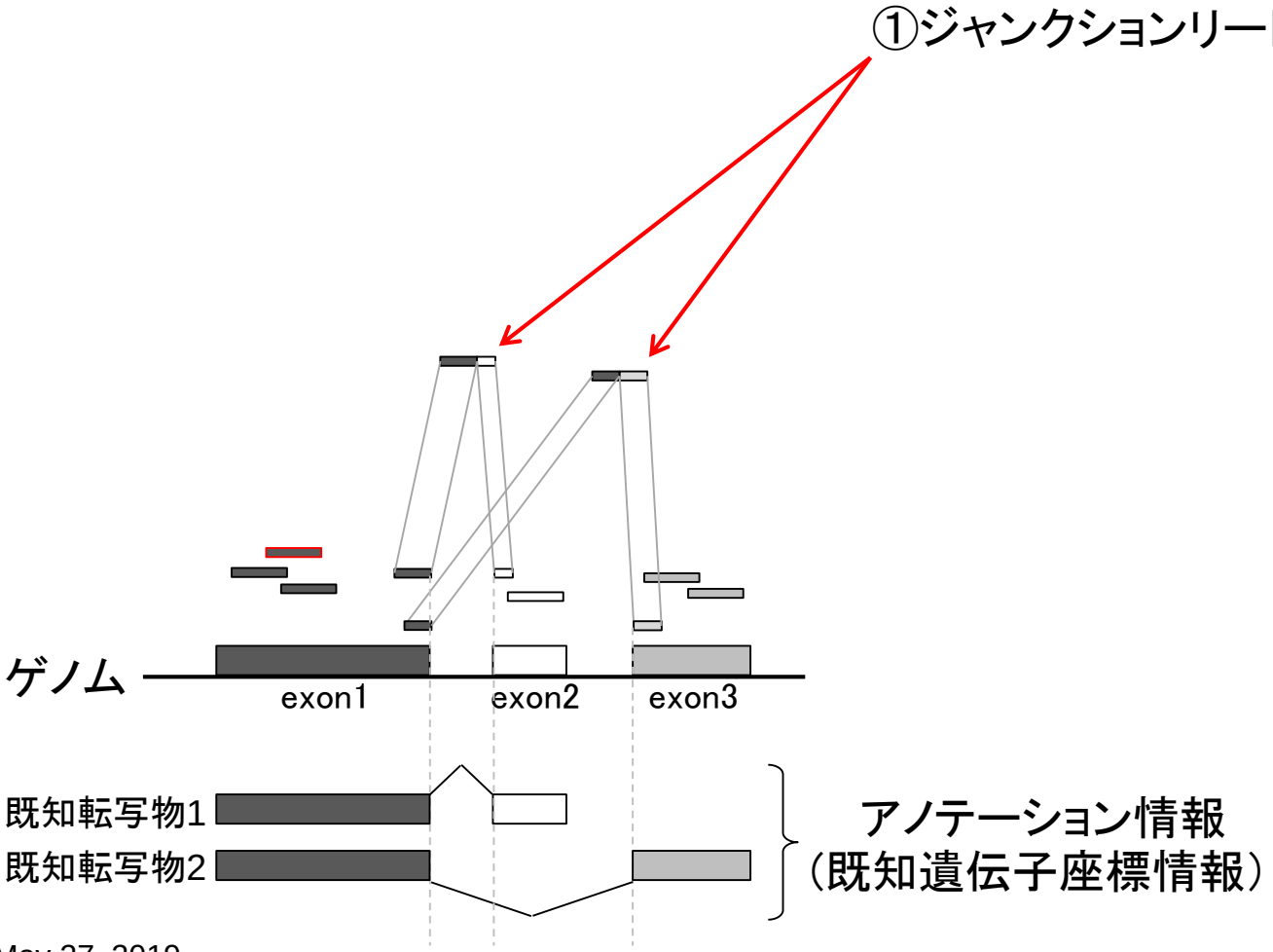


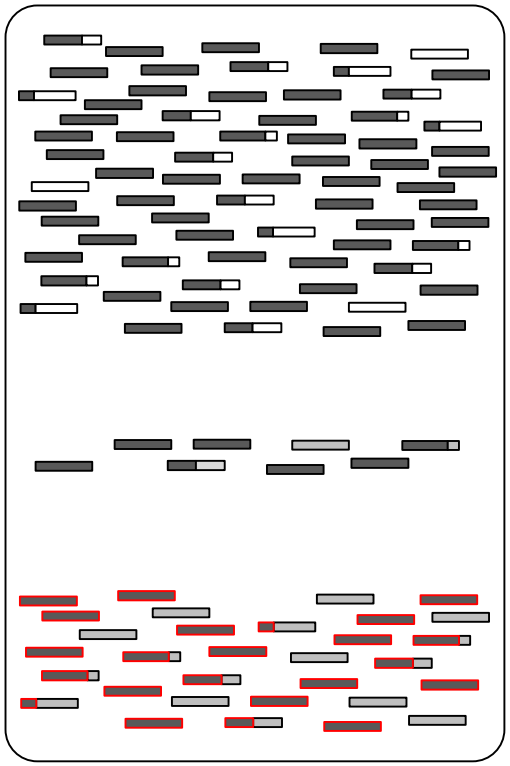


RNA-seqデータ

具体的な戦略

既知遺伝子（転写物）の座標情報と比較することで、答え合わせも可能

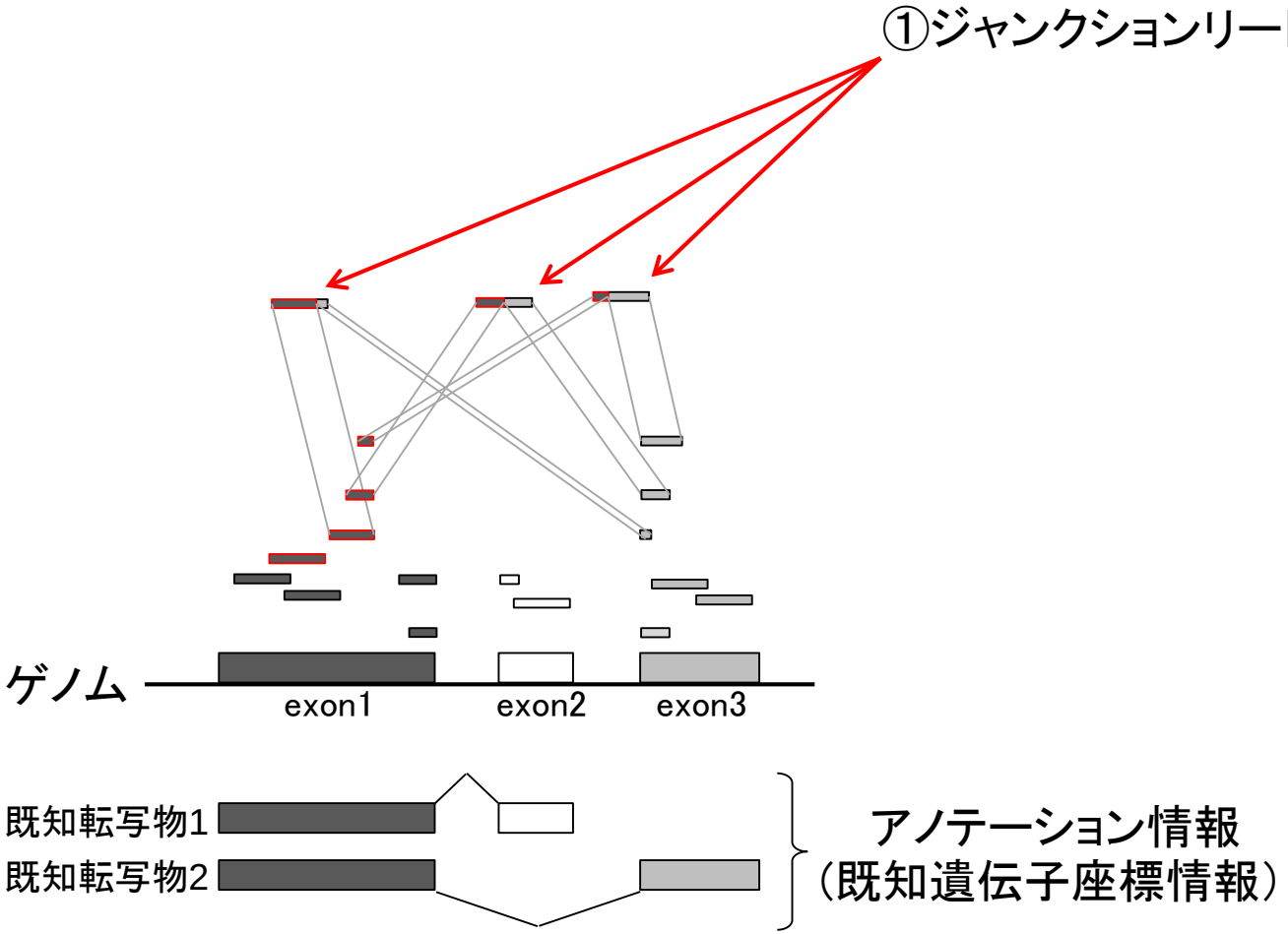


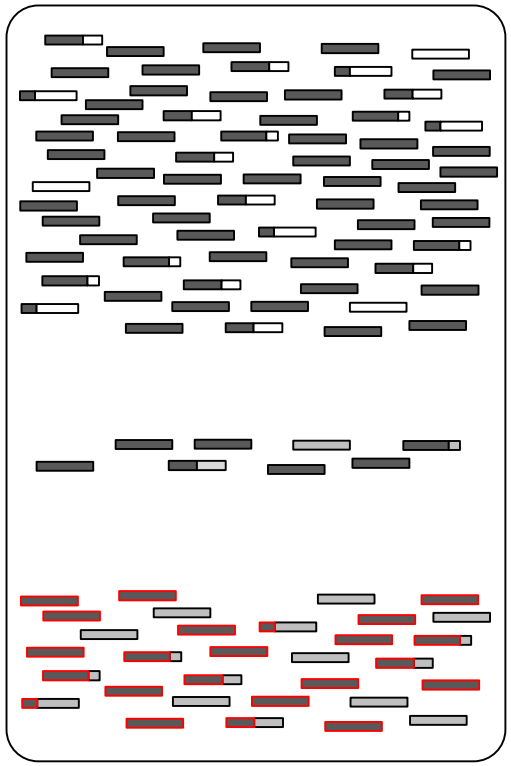


RNA-seqデータ

具体的な戦略

同様にして、他のジャンクションリードも既知転写物と比較することで…

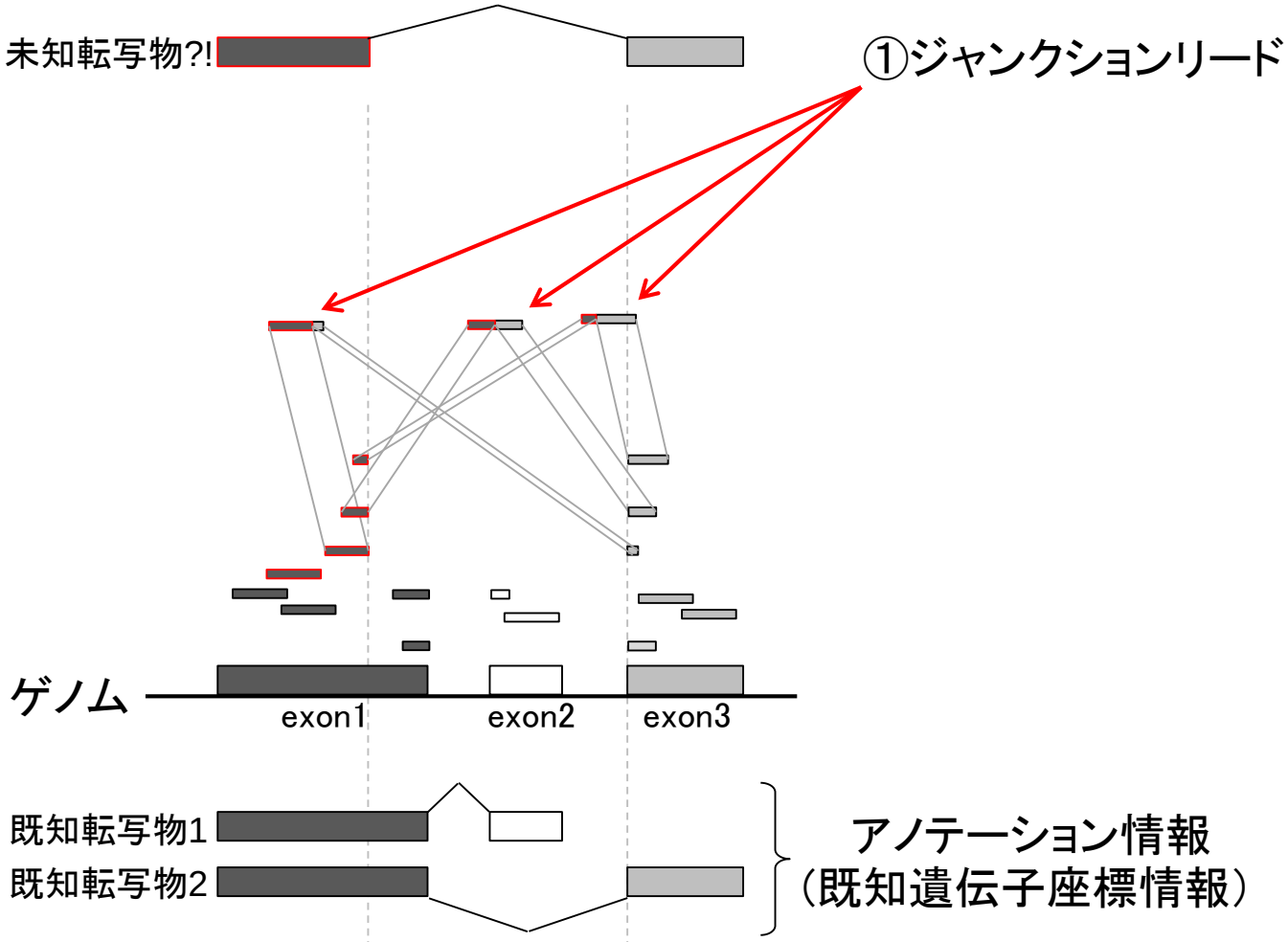


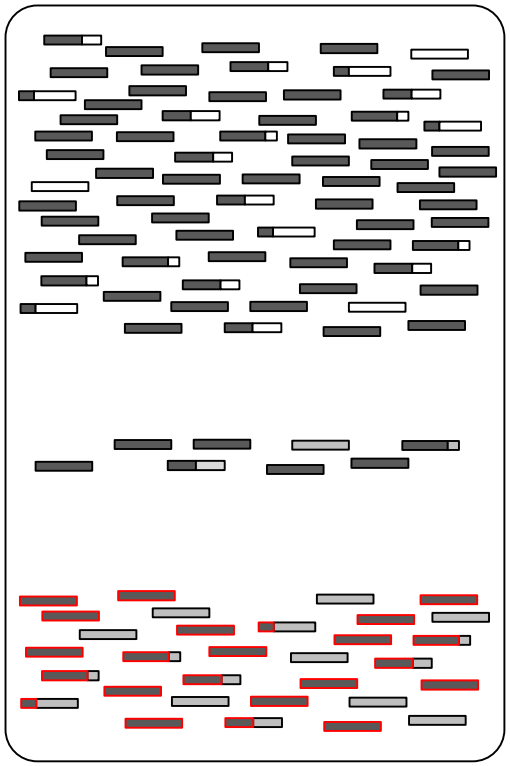


RNA-seqデータ

具体的な戦略

未知転写物(新規isoform)の同定も原理的に可能





RNA-seqデータ

新規転写物同定の例

RNA-seq(トランスクリプトーム解析)は、癌でよくみられる融合遺伝子の検出などにも利用されます。理由:そこそこ発現している転写物は原理的に検出可能だから。肺がんでみられるALK融合遺伝子(fusion gene)は有名な例ですが、それ以外の①新たな融合遺伝子の発見などに役立っています。主に「トランスクリプトーム**配列**解析」の話

https://www.ncbi.nlm.nih.gov/pubmed/25650807

NCBI Resources How To

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed Advanced

NCBI will be testing https on public web servers from 1:00-4:00 PM EDT (17:00-20:00 UTC) on Monday, October 24. You may experience problems with NCBI services, especially file downloads, during that time. Please plan accordingly. Read more.

Format: Abstract Send to

Genome Biol. 2015 Jan 5;16:7. doi: 10.1186/s13059-014-0558-0.

Identification of novel fusion genes in lung cancer using breakpoint assembly of transcriptome sequencing data.

①

Fernandez-Cuesta L, Sun R, Menon R, George J, Lorenz S, Meza-Zepeda LA, Peifer M, Plenker D, Heuckmann JM, Leenders F, Zander T, Dahmen I, Koker M, Schöttle J, Ullrich RT, Altmüller J, Becker C, Nürnberg P, Seidel H, Böhm D, Göke F, Ansén S, Russell PA, Wright GM, Wainer Z, Solomon B, Petersen I, Clement JH, Sänger J, Brustugun OT, Helland Å, Solberg S, Lund-Iversen M, Buettner R, Wolf J, Brambilla E, Vingron M, Perner S, Haas SA, Thomas RK.

Abstract

Genomic translocation events frequently underlie cancer development through generation of gene fusions with oncogenic properties. Identification of such fusion transcripts by transcriptome sequencing might help to discover new potential therapeutic targets. We developed TRUP (Tumor-specimen suited RNA-seq Unified Pipeline) (https://github.com/ruping/TRUP), a computational approach that combines split-read and read-pair analysis with de novo assembly for the identification of chimeric transcripts in cancer specimens. We apply TRUP to RNA-seq data of different tumor types, and find it to be more sensitive than alternative tools in detecting chimeric transcripts, such as secondary rearrangements in EML4-ALK-positive lung tumors, or recurrent inactivating rearrangements affecting RASSF8.

PMID: 25650807 PMCID: PMC4300615 DOI: 10.1186/s13059-014-0558-0

Full text links

Read free full text at BioMed Central

PMC FREE Full text

Save items

Add to Favorites

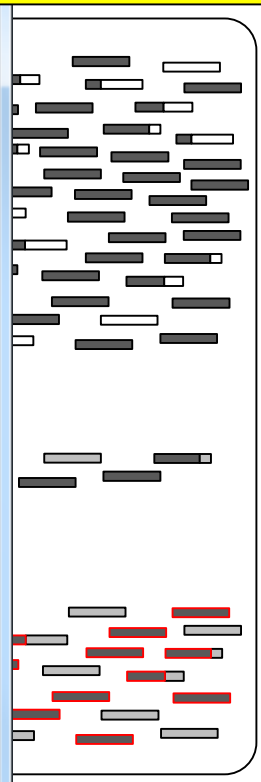
Similar articles

A transforming KIF5B and RET gene fusion in lung ac [Genome Res. 2012]

Sensitive and specific detection of EML4-ALK rearr [Lung Cancer. 2014]

Exon array profiling detects EML4-ALK fusion ir [Mol Cancer Res. 2009]

Review Sequential combination of



qデータ

Contents

- トランスクリプトーム解析技術の原理や特徴
  - RNA-seq（Illuminaの場合）、遺伝子≠転写物
- RNA-seqデータ解析のイメージ
  - マッピング→新規転写物の同定
- 様々な解析目的
  - 転写物配列取得、手法比較論文の紹介、ウェブツール
  - データ解析の全体像（入出力の関係や代表的なツール）
- アノテーションファイルの読み込みと課題1
  - Rで転写物配列取得のイントロ
- Rで転写物配列取得と課題2
  - アノテーションファイルとゲノム情報ファイルから
- 公共データベース
  - NGS全体（NCBI SRA, EMBL-EBI ENA, DDBJ SRA）
  - DRAの概要、クオリティスコアなど

様々な解析目的

Illumina社のNGS機器由来データを入力としたプログラムが多数派を占めている印象だが、①についてはロングリードデータの優位性も示されている。

- トランスクリプトーム配列解析 ①
  - ゲノム配列既知の場合：遺伝子構造推定、新規isoform同定など
  - ゲノム配列未知の場合：トランスクリプトーム用アセンブラを実行
  - 融合遺伝子同定に特化したプログラムもあり
  - 発現量推定まで行ってくれるものが多い
- トランスクリプトーム発現解析
  - 比較するサンプル間で発現変動している遺伝子または転写物の同定
  - 任意のリファレンス配列（ゲノムまたはトランスクリプトーム）にリードをマップし、カウントデータ取得、統計解析。ゲノム配列がなくてもトランスクリプトーム配列をアセンブリで取得すればリファレンスとして利用可能。

転写物配列取得

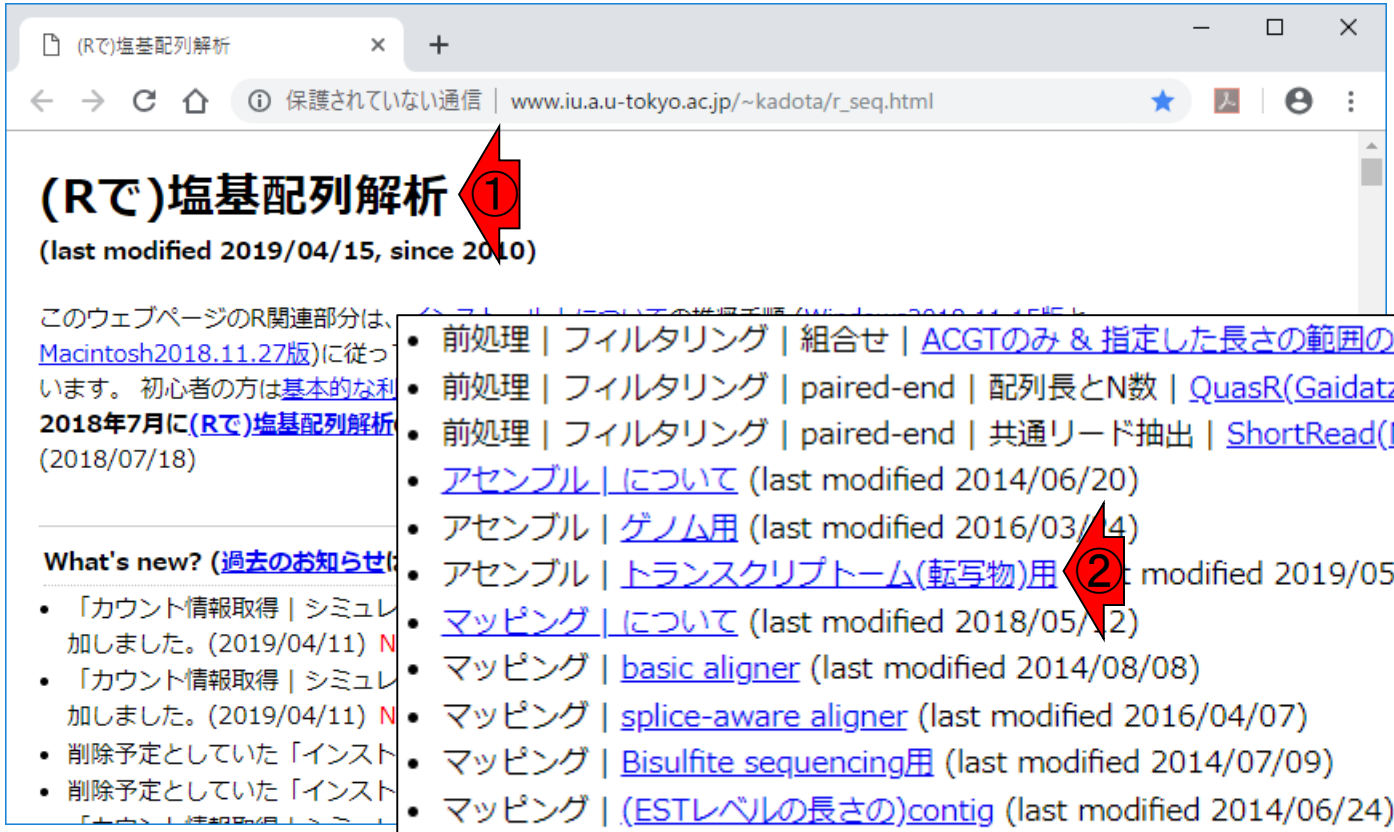
- ゲノム配列**既知**の場合

①の、②は遺伝子構造推定に相当する。つまり、「構造アノテーション」に相当するものであり、ゲノム配列上のどの領域に転写される領域が存在するのかという結果を返すものです。

転写物配列取得

- ゲノム配列**未知**の場合

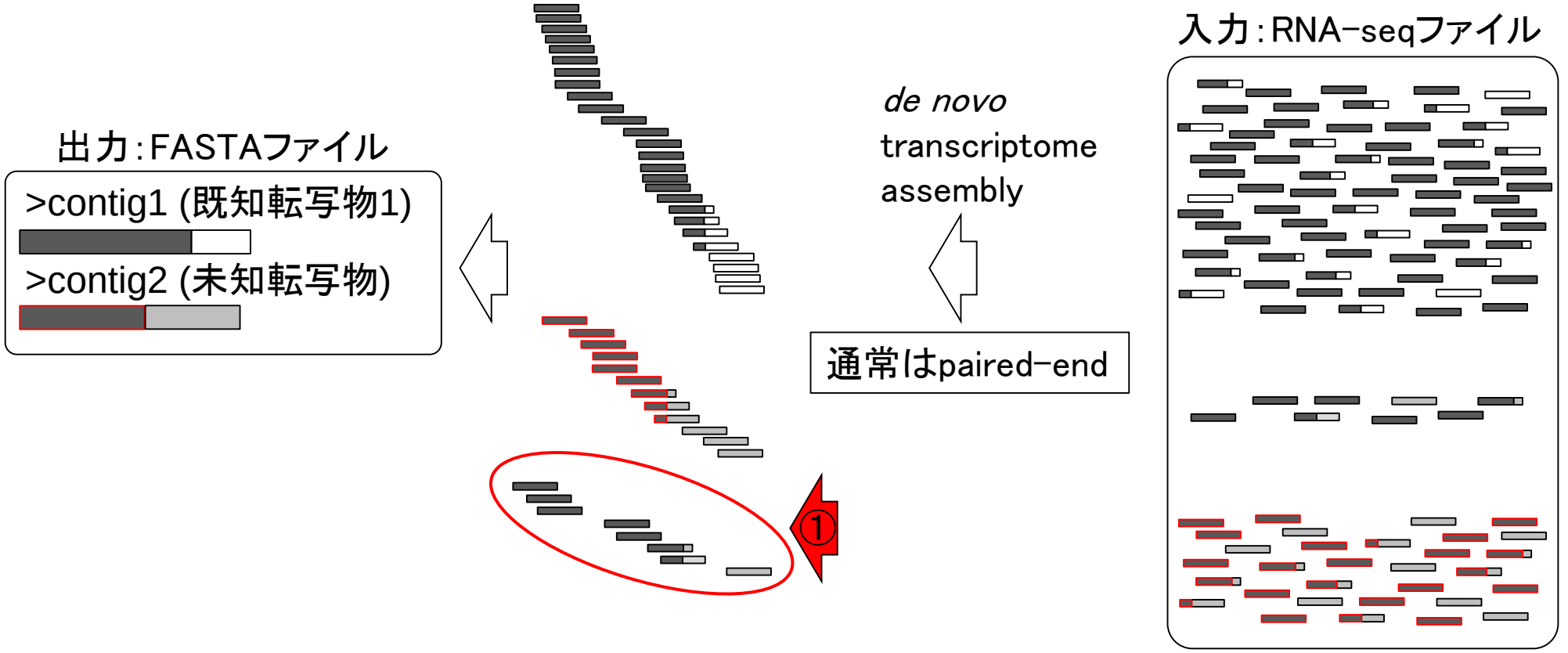
②トランスクリプトーム配列の*de novo*（1から、最初から、の意味）アセンブリに相当。多くのプログラムは発現量(FPKM値)も出力してくれます



転写物配列取得

- ゲノム配列未知の場合

*de novo*アセンブリのイメージ。①ターゲットサンプル中でそれほど発現していない転写物は、*de novo*アセンブリが原理的に困難。これはIllumina short-readデータをイメージしたもの



転写物配列取得

- ゲノム配列**未知**の場合

①のリンク先が、②です。バクテリア系はおそらく「ゲノム解読から遺伝子予測」の流れが多いと思います。*de novo*トランスクリプトームアセンブリのプログラムを利用するのは、基本的に「虫や植物系」かと思います。

- 前処理 | フィルタリング | 組合せ | ACGTのみ & 指定した長さの範囲の
- 前処理 | フィルタリング | paired-end | 配列長とN数 | QuasR(Gaidat
- 前処理 | フィルタリング | paired-end | 共通リード抽出 | ShortRead(
- アセンブル | について (last modified 2014/06/20)
- アセンブル | ゲノム用 (last modified 2016/03/14)
- アセンブル | トランスクリプトーム(転写物)用 ① t modified 2019/0
- マッピング | について (last modified 2018/05/12)
- マッピング | basic aligner (last modified 2014/08/08)
- マッピング | splice-aware aligner (last modified 2016/04/07)
- マッピング | Bisulfite sequencing用 (last modified 2014/07/09)
- マッピング | (ESTレベルの長さの)contig (last modified 2014/06/24
- マッピング | 基礎 (last modified 2013/06/19)
- マッピング | single-end | ゲノム | basic aligner(基礎) | QuasR(Gaic
- マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaic

②

アセンブル | トランスクリプトーム(転写物)用 NEW

Rパッケージはおそらくありません。

プログラム:

- Multiple-k : Surget-Groba and Montoya-Burgos, Genome Res., 2010
- Trans-ABySS : Robertson et al., Nat Methods, 2010
- Rnnotator : Martin et al., BMC Genomics, 2010
- Trinity : Grabherr et al., Nat Biotechnol, 2011
- Oases : Schulz et al., Bioinformatics, 2012
- EBARDenovo : Chu et al., Bioinformatics, 2013
- BRANCH : Bao et al., Bioinformatics, 2013
- IDBA-tran : Peng et al., Bioinformatics, 2013
- SOAPdenovo-Trans : Xie et al., Bioinformatics, 2014
- VTBuilder : Archer et al., BMC Bioinformatics, 2014
- Rockhopper2(バクテリア用) : Tjaden B, Genome Biol., 2015
- DETONATE(RSEM-EVAL) : Li et al., Genome Biol., 2014
- Bridger : Chang et al., Genome Biol., 2015
- IFRAT : Mbandi et al., BMC Bioinformatics, 2015
- SCERNA(主に植物) : Honaas et al., PLoS One, 2016
- BinPacker : Liu et al., PLoS Comput Biol., 2016
- TraRECo(Windows/Linux用) : Yoon et al., BMC Genomics, 2018

Review、ガイドライン、パイプライン系:

- Review : Martin and Wang, Nat Rev Genet., 2011
- ガイドライン : Haznedaroglu et al., BMC Bioinformatics, 2012
- Review : Góngora-Castillo, Nat Prod Rep., 2013
- ガイドライン : Yang and Smith, BMC Genomics, 2013
- ガイドライン : O'Neil et al., BMC Genomics, 2013
- ガイドライン : Feldmesser et al., BMC Genomics, 2014
- パイプライン(454用) : Melicher et al., BMC Genomics, 2014
- パイプライン(組合せ系; SAMP and CDTA) : He et al., BMC Genomics, 2015
- 手法比較(Bridgerがよかった) : Rana et al., PLoS One, 2016
- ガイドライン系(multiple-k strategyのどのあたりまでk値を試すかに関する議論) : Durai and Schulz, Bioinformatics, 2016
- 手法比較(リファレンス配列があってもde novoをやる価値はあるとのこと) : Wang and Gribskov, Bioinformatics, 2017
- 手法比較 : Hölzer and Marz, Gigascience, 2019

転写物配列取得

一般にどのような評価基準でプログラムの良しあしを判断しているのか？自分が取り扱う生物種でよいパフォーマンスを示すことが期待されるプログラムはどれか？マニュアルが丁寧で実用的なものはどれか？が気になるところです。

アセンブル | トランスクリプトーム(転写物)用 NEW

Rパッケージはおそらくありません。

プログラム:

- Multiple-k : Surget-Groba and Montoya-Burgos, Genome Res., 2010
- Trans-ABySS : Robertson et al., Nat Methods, 2010
- Rnnotator : Martin et al., BMC Genomics, 2010
- Trinity : Grabherr et al., Nat Biotechnol, 2011
- Oases : Schulz et al., Bioinformatics, 2012
- EBARDenovo : Chu et al., Bioinformatics, 2013
- BRANCH : Bao et al., Bioinformatics, 2013
- IDBA-tran : Peng et al., Bioinformatics, 2013
- SOAPdenovo-Trans : Xie et al., Bioinformatics, 2014
- VTBuilder : Archer et al., BMC Bioinformatics, 2014
- Rockhopper2(バクテリア用) : Tjaden B, Genome Biol., 2015
- DETONATE(RSEM-EVAL) : Li et al., Genome Biol., 2014
- Bridger : Chang et al., Genome Biol., 2015
- IFRAT : Mbandi et al., BMC Bioinformatics, 2015
- SCERNA(主に植物) : Honaas et al., PLoS One, 2016
- BinPacker : Liu et al., PLoS Comput Biol., 2016
- TraRECo(Windows/Linux用) : Yoon et al., BMC Genomics, 2018

Review、ガイドライン、パイプライン系:

- Review : Martin and Wang, Nat Rev Genet., 2011
- ガイドライン : Haznedaroglu et al., BMC Bioinformatics, 2012
- Review : Góngora-Castillo, Nat Prod Rep., 2013
- ガイドライン : Yang and Smith, BMC Genomics, 2013
- ガイドライン : O'Neil et al., BMC Genomics, 2013
- ガイドライン : Feldmesser et al., BMC Genomics, 2014
- パイプライン(454用) : Melicher et al., BMC Genomics, 2014
- パイプライン(組合せ系; SAMP and CDTA) : He et al., BMC Genomics, 2015
- 手法比較(Bridgerがよかった) : Rana et al., PLoS One, 2016
- ガイドライン系(multiple-k strategyのどのあたりまでk値を試すかに関する議論) : Durai and Schulz, Bioinformatics, 2016
- 手法比較(リファレンス配列があってもde novoをやる価値はあるとのこと) : Wang and Gribskov, Bioinformatics, 2017
- 手法比較 : Hölzer and Marz, Gigascience, 2019

転写物配列取得

一般にどのような評価基準でプログラムの良しあしを判断しているのか？自分が取り扱う生物種でよいパフォーマンスを示すことが期待されるプログラムはどれか？マニュアルが丁寧で実用的なものはどれか？が気になるところです。ググるのも一手ですが、①Reviewや手法比較系論文の、②最新のものを眺めるとよいと思います。

アセンブル | トランスクリプトーム(転写物)用 NEW

Rパッケージはおそらくありません。

プログラム:

- Multiple-k : Surget-Groba and Montoya-Burgos, Genome Res., 2010
- Trans-ABySS : Robertson et al., Nat Methods, 2010
- Rnnotator : Martin et al., BMC Genomics, 2010
- Trinity : Grabherr et al., Nat Biotechnol, 2011
- Oases : Schulz et al., Bioinformatics, 2012
- EBARDenovo : Chu et al., Bioinformatics, 2013
- BRANCH : Bao et al., Bioinformatics, 2013
- IDBA-tran : Peng et al., Bioinformatics, 2013
- SOAPdenovo-Trans : Xie et al., Bioinformatics, 2014
- VTBuilder : Archer et al., BMC Bioinformatics, 2014
- Rockhopper2(バクテリア用) : Tjaden B, Genome Biol., 2015
- DETONATE(RSEM-EVAL) : Li et al., Genome Biol., 2014
- Bridger : Chang et al., Genome Biol., 2015
- IFRAT : Mbandi et al., BMC Bioinformatics, 2015
- SCERNA(主に植物) : Honaas et al., PLoS One, 2016
- BinPacker : Liu et al., PLoS Comput Biol., 2016
- TraRECo(Windows/Linux用) : Yoon et al., BMC Genomics, 2018

Review、ガイドライン、パイプライン系: ①

- Review : Martin and Wang, Nat Rev Genet., 2011
- ガイドライン : Haznedaroglu et al., BMC Bioinformatics, 2012
- Review : Góngora-Castillo, Nat Prod Rep., 2013
- ガイドライン : Yang and Smith, BMC Genomics, 2013
- ガイドライン : O'Neil et al., BMC Genomics, 2013
- ガイドライン : Feldmesser et al., BMC Genomics, 2014
- パイプライン(454用) : Melicher et al., BMC Genomics, 2014
- パイプライン(組合せ系; SAMP and CDTA) : He et al., BMC Genomics, 2015
- 手法比較(Bridgerがよかった) : Rana et al., PLoS One, 2016
- ガイドライン系(multiple-k strategyのどのあたりまで値を試すかに関する議論) : Durai and Schulz, Bioinformatics, 2016
- 手法比較(リファレンス配列があってもde novoをやる意義はあるとのこと) : Wang and Gribskov, Bioinformatics, 2017
- 手法比較 : Hölzer and Marz, Gigascience, 2019 ②

Contents

- トランスクリプトーム解析技術の原理や特徴
  - RNA-seq（Illuminaの場合）、遺伝子≠転写物
- RNA-seqデータ解析のイメージ
  - マッピング→新規転写物の同定
- 様々な解析目的
  - 転写物配列取得、手法比較論文の紹介、ウェブツール
  - データ解析の全体像（入出力の関係や代表的なツール）
- アノテーションファイルの読み込みと課題1
  - Rで転写物配列取得のイントロ
- Rで転写物配列取得と課題2
  - アノテーションファイルとゲノム情報ファイルから
- 公共データベース
  - NGS全体（NCBI SRA, EMBL-EBI ENA, DDBJ SRA）
  - DRAの概要、クオリティスコアなど

手法比較論文

②2019年5月にpublishされた手法比較論文は、
③様々な生物種について評価を行っており、丁寧によくまとめられている。但し、④short-read用。

Gigascience. 2019 May 1;8(5). pii: giz039. doi: 10.1093/gigascience/giz039.

De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers.

Hölzer M[1,2], Marz M[1,2,3].

Author information

Abstract

BACKGROUND: In recent years, massively parallel complementary DNA sequencing (RNA sequencing [RNA-Seq]) has emerged as a fast, cost-effective, and robust technology to study entire transcriptomes in various manners. In particular, for non-model organisms and in the absence of an appropriate reference genome, RNA-Seq is used to reconstruct the transcriptome de novo. Although the de novo transcriptome assembly of non-model organisms has been on the rise recently and new tools are frequently developing, there is still a knowledge gap about which assembly software should be used to build a comprehensive de novo assembly.

RESULTS: Here, we present a large-scale comparative study in which 10 de novo assembly tools are applied to 9 RNA-Seq data sets spanning different kingdoms of life. Overall, we built >200 single assemblies and evaluated their performance on a combination of 20 biological-based and reference-free metrics. Our study is accompanied by a comprehensive and extensible Electronic Supplement that summarizes all data sets, assembly execution instructions, and evaluation results. Trinity, SPAdes, and Trans-ABySS, followed by Bridger and SOAPdenovo-Trans, generally outperformed the other tools compared. Moreover, we observed species-specific differences in the performance of each assembler. No tool delivered the best results for all data sets.

CONCLUSIONS: We recommend a careful choice and normalization of evaluation metrics to select the best assembling results as a critical step in the reconstruction of a comprehensive de novo transcriptome assembly.

手法比較論文

結論としては、⑤の記述から、昔からよく利用されている⑥Trinityというプログラムでよいのだろうと判断。

Gigascience. 2019 May 1;8(5). pii: giz039. doi: 10.1093/gigascience/giz039.

De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers.

Hölzer M[1,2], Marz M[1,2,3].

⊕ Author information

Abstract

BACKGROUND: In recent years, massively parallel complementary DNA sequencing (RNA sequencing [RNA-Seq]) has emerged as a fast, cost-effective, and robust technology to study entire transcriptomes in various manners. In particular, for non-model organisms and in the absence of an appropriate reference genome, RNA-Seq is used to reconstruct the transcriptome de novo. Although the de novo transcriptome assembly of non-model organisms has been on the rise recently and new tools are frequently developing, there is still a knowledge gap about which assembly software should be used to build a comprehensive de novo assembly.

RESULTS: Here, we present a large-scale comparative study in which 10 de novo assembly tools are applied to 9 RNA-Seq data sets spanning different kingdoms of life. Overall, we built >200 single assemblies and evaluated their performance on a combination of 20 biological-based and reference-free metrics. Our study is accompanied by a comprehensive and extensible Electronic Supplement that summarizes all data sets, assembly execution instructions, and evaluation results. Trinity, SPAdes, and Trans-ABySS, followed by Bridger and SOAPdenovo-Trans, generally outperformed the other tools compared. Moreover, we observed species-specific differences in the performance of each assembler. No tool delivered the best results for all data sets.

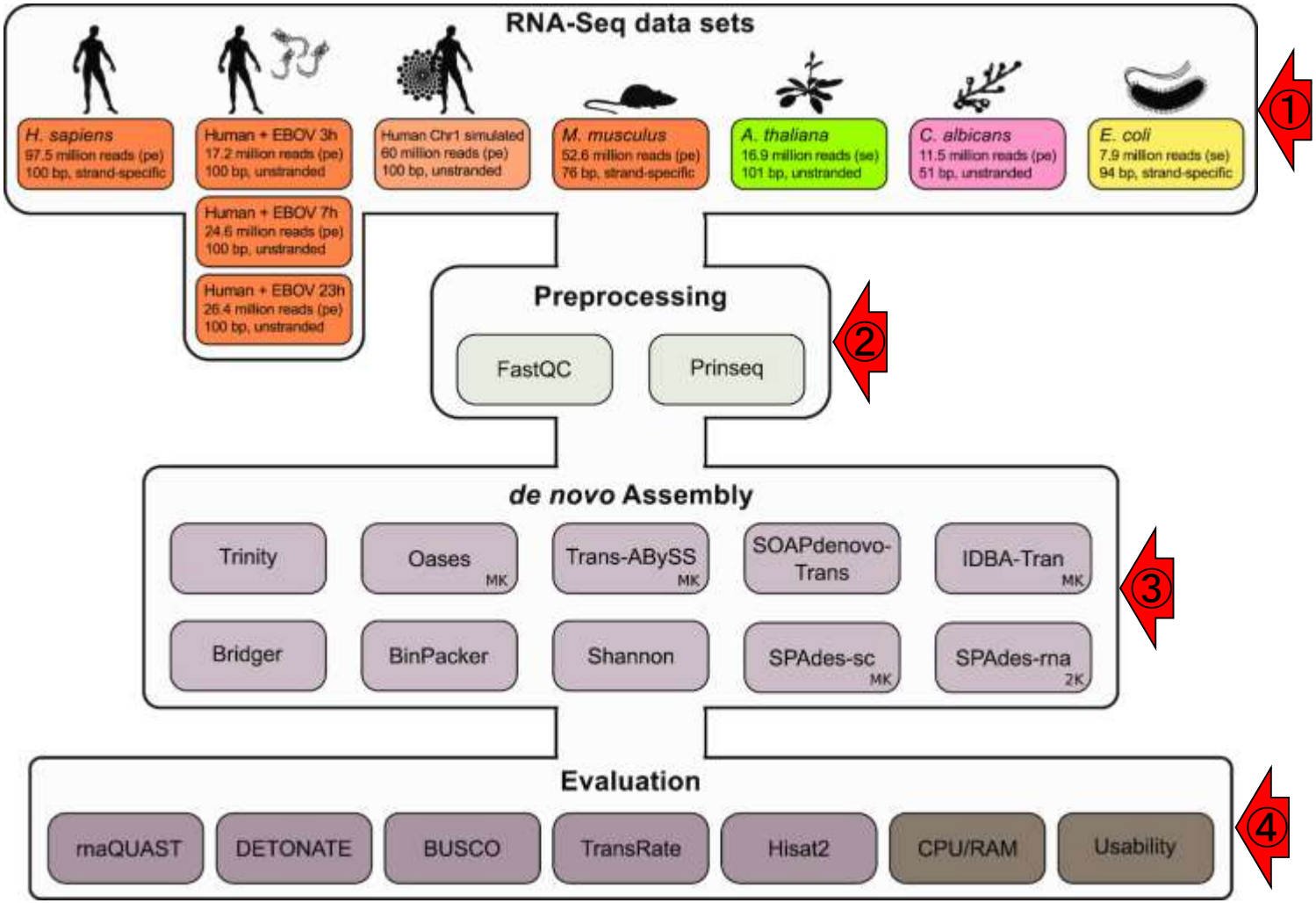
CONCLUSIONS: We recommend a careful choice and normalization of evaluation metrics to select the best assembling results as a critical step in the reconstruction of a comprehensive de novo transcriptome assembly.

⑥

⑤

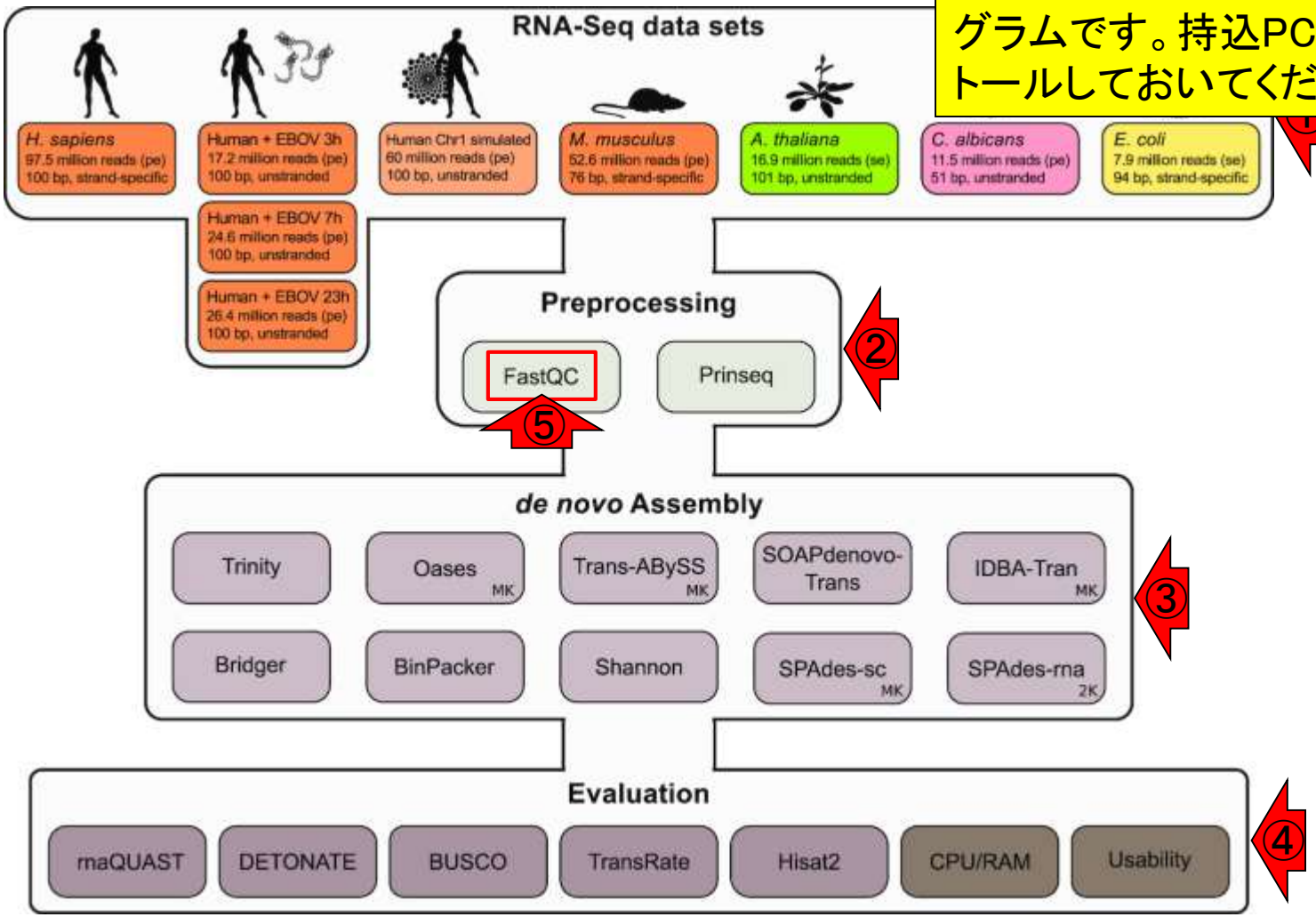
手法比較論文の図1

この論文で使われた、①データセット、②前処理、③アセンブリプログラム、④評価指標



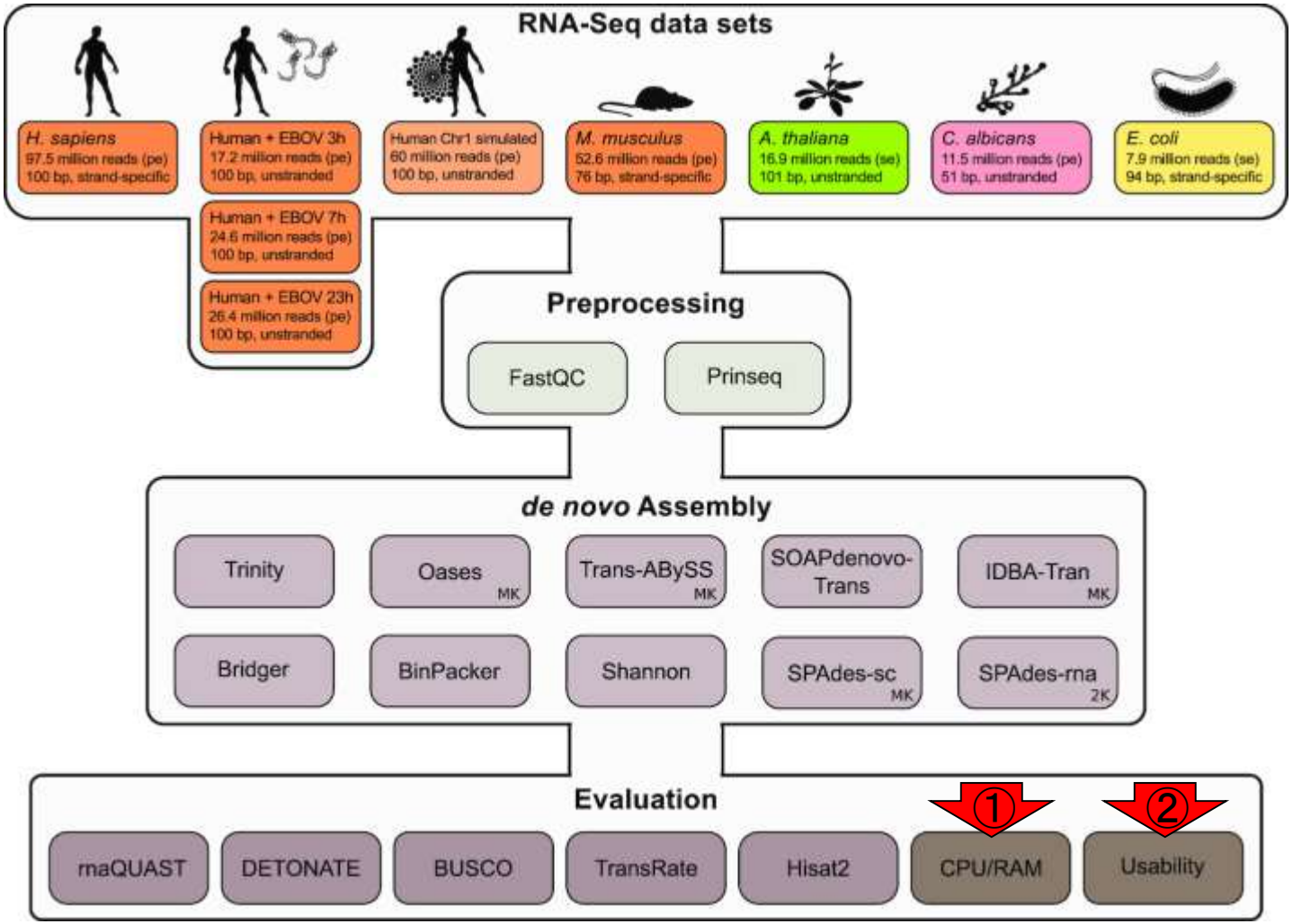
手法比較論文の図1

この論文で使われた、①データセット、②前処理、③アセンブリプログラム、④評価指標。⑤FastQCは来週利用します。FastQCはJavaプログラムです。持込PCのヒトは、予めJavaをインストールしておいてください。



手法比較論文の図1

評価指標として、①計算機資源や、②使いやすさが含まれている。これは現実問題として重要。



手法比較論文の表3

Table 3です。①アセンブルプログラム(Assembler)。②インストールが簡単かどうか。

①

②

Table 3 Overview of the different *de novo* assembly tools evaluated in this study

Assembler	Version	MK	Setup	Usage	Runtime			Memory (GB)			Source	Year
					Min	Max	Median	Min	Max	Median		
Trans-ABySS	2.0.1	Yes	🙂	🙂	16 m	2 d 6 h 23 m	11 h 11 m	0.6	49.2	19.7	[9]	2010
Trinity	2.8.4	No	🙂	😄	28 m	1 d 20 h 10 m	6 h 40 m	7.2	243.9	27.7	[10]	2011
Oases[a]	0.2.08	Yes	🙂	🙂	25 m	8 d 15 h 45 m	6 h 47 m	3.1	110.2	31.3	[11]	2012
SPAdes-sc[b]	3.13.0	Yes	😄	😄	16 m	7 h 52 m	2 h 26 m	5.0	37.4	25.3	[18]	2012
SPAdes-rna[b]	3.13.0	Yes[c]	😄	😄	11 m	7 h 24 m	2 h 17 m	5.0	44.2	19.5	[17]	2018
IDBA-Tran	1.1.1	Yes	🙂	🙂	7 m	8 h 49 m	2 h 44 m	0.6	29.1	9.6	[12]	2013
SOAPdenovo-Trans	1.03	No	🙂	😕	1 m	1 h 48 m	24 m	2.1	45.6	26.4	[13]	2014
Bridger[d]	14-12-01	No	🙂	😕	11 m	21 h 11 m	5 h 9 m	1.6	109.3	30.4	[14]	2015
BinPacker[d]	1.0	No	🙂	😄	5 m	15 h 57 m	3 h 3 m	1.5	96.2	27.9	[15]	2016
Shannon	0.0.2	No	😕	🙂	9 m	10 h 45 m	3 h 18 m	3.8	121.4	83.6	[16]	2016

We rated our experiences regarding the installation and usability of each tool (😄: excellent; 🙂: good; 😕: unsatisfactory). These experiences might be subjective; nevertheless, we want to share them to give non-experienced users an idea of how difficult it is to get each tool installed (Setup) and executed (Usage) (see Methods for details). For Trinity, we observed high memory peaks at the beginning of the calculations for large (human, mouse) data sets, which immediately returned to moderate memory levels after a few minutes. More details about runtime and memory consumption can be found in Electronic Supplement Fig. S11. MK: presence of a built-in multiple k-mer approach and the ability to automatically integrate the output of different k-mer runs.

[a] Oases was used on top of the *de novo* genome assembler Velvet (v1.2.10) [45].

[b] SPAdes, originally designed as a *de novo* genome assembler for single-cell data, was used in single-cell modus (--sc) and RNA-Seq modus (--rna).

[c] When running SPAdes in RNA-Seq modus, 2 k-mer values are used by default.

[d] Bridger and BinPacker are based on a splicing graph construction instead of de Bruijn graphs.

手法比較論文の表3

③実行に要する時間。第3回で行う予定の発現変動解析のように一瞬で計算が終わるわけではないことがわかる。

Table 3: Overview of the different *de novo* assembly tools evaluated in this study

Assembler	Version	MK	Setup	Usage	Runtime			Memory (GB)			Source	Year
					Min	Max	Median	Min	Max	Median		
Trans-ABySS	2.0.1	Yes	🙂	🙂	16 m	2 d 6 h 23 m	11 h 11 m	0.6	49.2	19.7	[9]	2010
Trinity	2.8.4	No	🙂	😄	28 m	1 d 20 h 10 m	6 h 40 m	7.2	243.9	27.7	[10]	2011
Oases[a]	0.2.08	Yes	🙂	🙂	25 m	8 d 15 h 45 m	6 h 47 m	3.1	110.2	31.3	[11]	2012
SPAdes-sc[b]	3.13.0	Yes	😄	😄	16 m	7 h 52 m	2 h 26 m	5.0	37.4	25.3	[18]	2012
SPAdes-rna[b]	3.13.0	Yes[c]	😄	😄	11 m	7 h 24 m	2 h 17 m	5.0	44.2	19.5	[17]	2018
IDBA-Tran	1.1.1	Yes	🙂	🙂	7 m	8 h 49 m	2 h 44 m	0.6	29.1	9.6	[12]	2013
SOAPdenovo-Trans	1.03	No	🙂	😕	1 m	1 h 48 m	24 m	2.1	45.6	26.4	[13]	2014
Bridger[d]	14-12-01	No	🙂	😕	11 m	21 h 11 m	5 h 9 m	1.6	109.3	30.4	[14]	2015
BinPacker[d]	1.0	No	🙂	😄	5 m	15 h 57 m	3 h 3 m	1.5	96.2	27.9	[15]	2016
Shannon	0.0.2	No	😕	🙂	9 m	10 h 45 m	3 h 18 m	3.8	121.4	83.6	[16]	2016

We rated our experiences regarding the installation and usability of each tool (😄: excellent; 🙂: good; 😕: unsatisfactory). These experiences might be subjective; nevertheless, we want to share them to give non-experienced users an idea of how difficult it is to get each tool installed (Setup) and executed (Usage) (see Methods for details). For Trinity, we observed high memory peaks at the beginning of the calculations for large (human, mouse) data sets, which immediately returned to moderate memory levels after a few minutes. More details about runtime and memory consumption can be found in Electronic Supplement Fig. S11. MK: presence of a built-in multiple k-mer approach and the ability to automatically integrate the output of different k-mer runs.

[a] Oases was used on top of the *de novo* genome assembler Velvet (v1.2.10) [45].

[b] SPAdes, originally designed as a *de novo* genome assembler for single-cell data, was used in single-cell modus (--sc) and RNA-Seq modus (--rna).

[c] When running SPAdes in RNA-Seq modus, 2 k-mer values are used by default.

[d] Bridger and BinPacker are based on a splicing graph construction instead of de Bruijn graphs.

手法比較論文の表3

④実行に要するメモリ容量。アセンブリ系はノートPCレベルでは実質的に実行不能であることを正しく認識すべし。

Table 3: Overview of the different *de novo* assembly tools evaluated in this study

④

Assembler	Version	MK	Setup	Usage	Runtime			Memory (GB)			Source	Year
					Min	Max	Median	Min	Max	Median		
Trans-ABySS	2.0.1	Yes	🙂	🙂	16 m	2 d 6 h 23 m	11 h 11 m	0.6	49.2	19.7	[9]	2010
Trinity	2.8.4	No	🙂	😄	28 m	1 d 20 h 10 m	6 h 40 m	7.2	243.9	27.7	[10]	2011
Oases[a]	0.2.08	Yes	🙂	🙂	25 m	8 d 15 h 45 m	6 h 47 m	3.1	110.2	31.3	[11]	2012
SPAdes-sc[b]	3.13.0	Yes	😄	😄	16 m	7 h 52 m	2 h 26 m	5.0	37.4	25.3	[18]	2012
SPAdes-rna[b]	3.13.0	Yes[c]	😄	😄	11 m	7 h 24 m	2 h 17 m	5.0	44.2	19.5	[17]	2018
IDBA-Tran	1.1.1	Yes	🙂	🙂	7 m	8 h 49 m	2 h 44 m	0.6	29.1	9.6	[12]	2013
SOAPdenovo-Trans	1.03	No	🙂	😕	1 m	1 h 48 m	24 m	2.1	45.6	26.4	[13]	2014
Bridger[d]	14-12-01	No	🙂	😕	11 m	21 h 11 m	5 h 9 m	1.6	109.3	30.4	[14]	2015
BinPacker[d]	1.0	No	🙂	😄	5 m	15 h 57 m	3 h 3 m	1.5	96.2	27.9	[15]	2016
Shannon	0.0.2	No	😕	🙂	9 m	10 h 45 m	3 h 18 m	3.8	121.4	83.6	[16]	2016

We rated our experiences regarding the installation and usability of each tool (😄: excellent; 🙂: good; 😕: unsatisfactory). These experiences might be subjective; nevertheless, we want to share them to give non-experienced users an idea of how difficult it is to get each tool installed (Setup) and executed (Usage) (see Methods for details). For Trinity, we observed high memory peaks at the beginning of the calculations for large (human, mouse) data sets, which immediately returned to moderate memory levels after a few minutes. More details about runtime and memory consumption can be found in Electronic Supplement Fig. S11. MK: presence of a built-in multiple k-mer approach and the ability to automatically integrate the output of different k-mer runs.

[a] Oases was used on top of the *de novo* genome assembler Velvet (v1.2.10) [45].

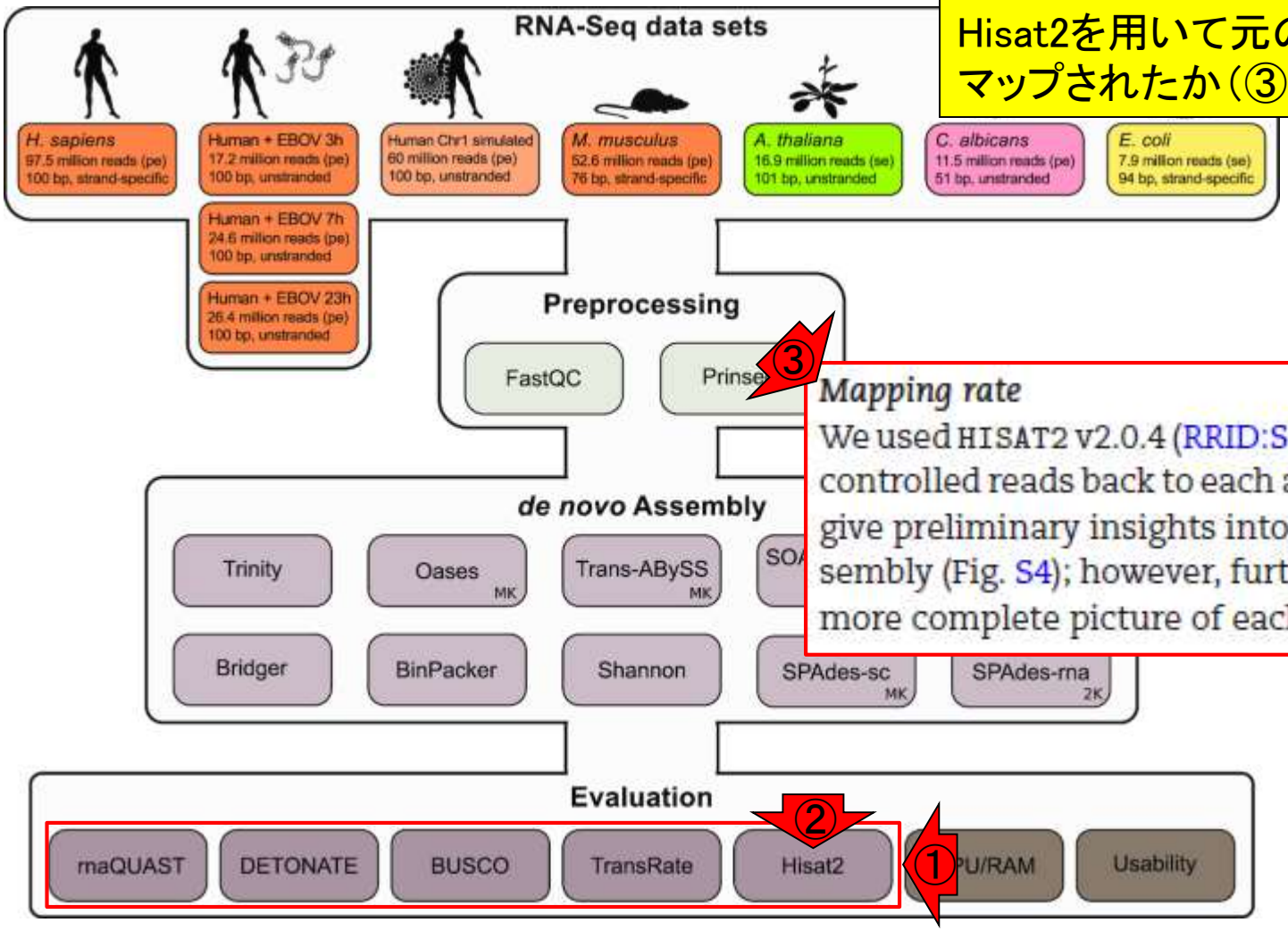
[b] SPAdes, originally designed as a *de novo* genome assembler for single-cell data, was used in single-cell modus (--sc) and RNA-Seq modus (--rna).

[c] When running SPAdes in RNA-Seq modus, 2 k-mer values are used by default.

[d] Bridger and BinPacker are based on a splicing graph construction instead of de Bruijn graphs.

手法比較論文の図1

①他の評価指標。これにはプログラム名も含まれる。例えば、②Hisat2はマッピングプログラム。アセンブルで得られた転写物配列群に対してHisat2を用いて元のリードをマップし、どの程度マップされたか（③マップ率）を調べている。

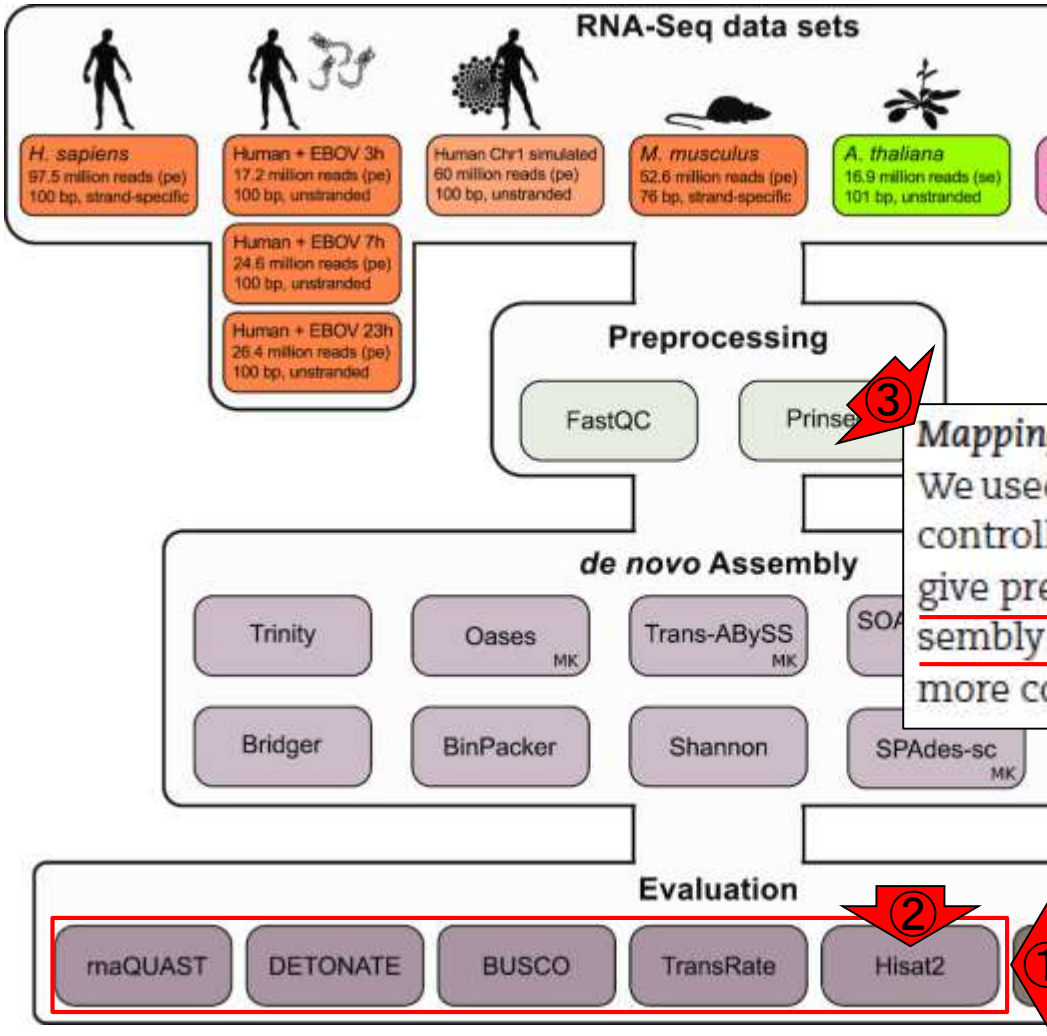


Mapping rate

We used HISAT2 v2.0.4 (RRID:SCR_015530) [41] to map the quality-controlled reads back to each assembly. The re-mapping rate can give preliminary insights into the quality of a transcriptome assembly (Fig. S4); however, further metrics are needed to assess a more complete picture of each assembler's performance.

手法比較論文の図1

①他の評価指標。これにはプログラム名も含まれる。例えば、②Hisat2はマッピングプログラム。アセンブルで得られた転写物配列群に対してHisat2を用いて元のリードをマップし、どの程度マップされたか（③マップ率）を調べている。④マップ率が低ければ、アセンブリに使われていないリードが沢山あることを意味する。



③

Mapping rate

We used HISAT2 v2.0.4 (RRID:SCR_015[illegible]) [41] to map the quality-controlled reads back to each assembly. The re-mapping rate can give preliminary insights into the quality of a transcriptome assembly (Fig. S4); however, further metrics are needed to assess a more complete picture of each assembler's performance.

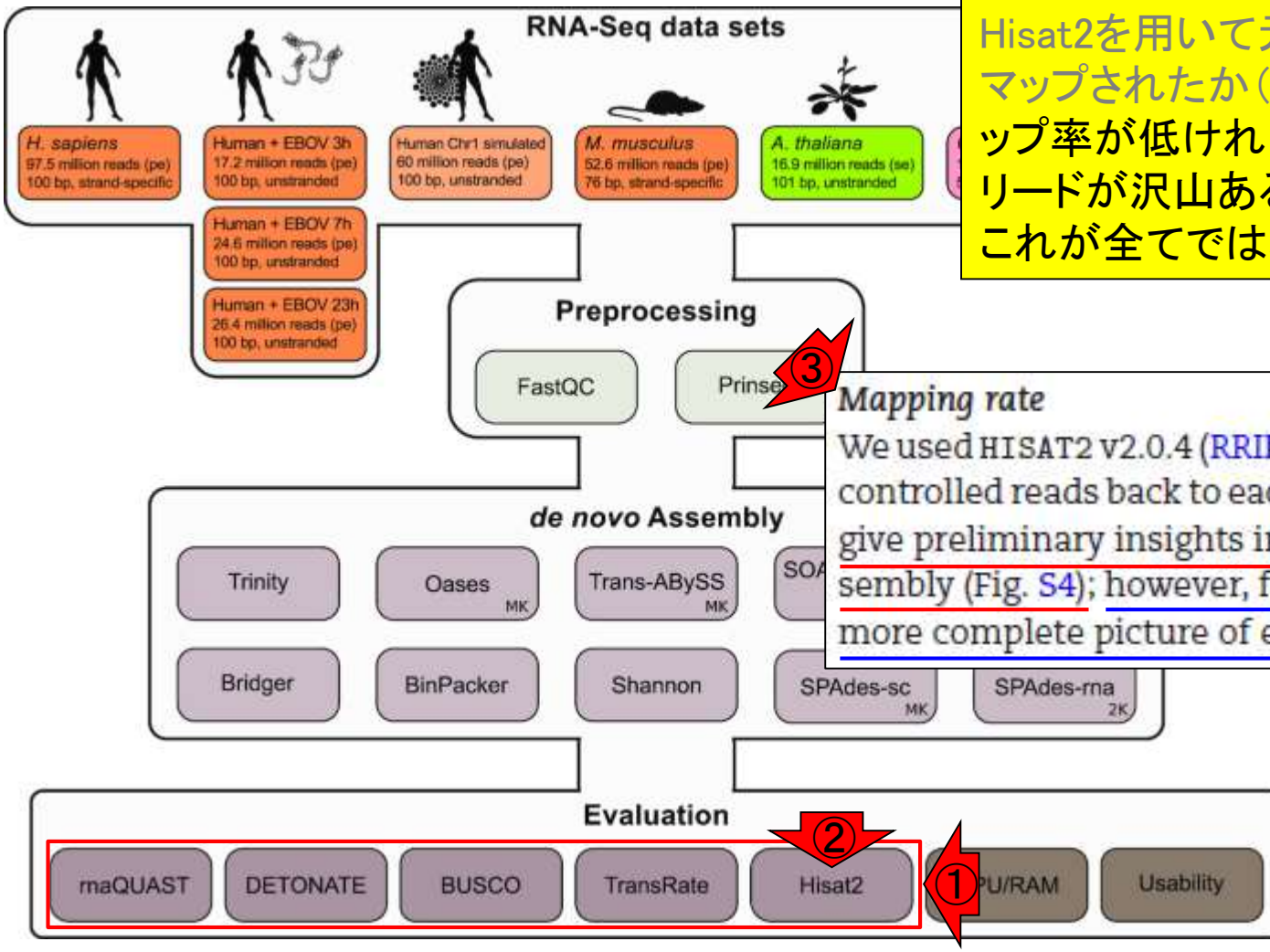
④

②

①

手法比較論文の図1

①他の評価指標。これにはプログラム名も含まれる。例えば、②Hisat2はマッピングプログラム。アセンブルで得られた転写物配列群に対してHisat2を用いて元のリードをマップし、どの程度マップされたか（③マップ率）を調べている。④マップ率が低ければ、アセンブリに使われていないリードが沢山あることを意味するが、⑤もちろんこれが全てではない。



③

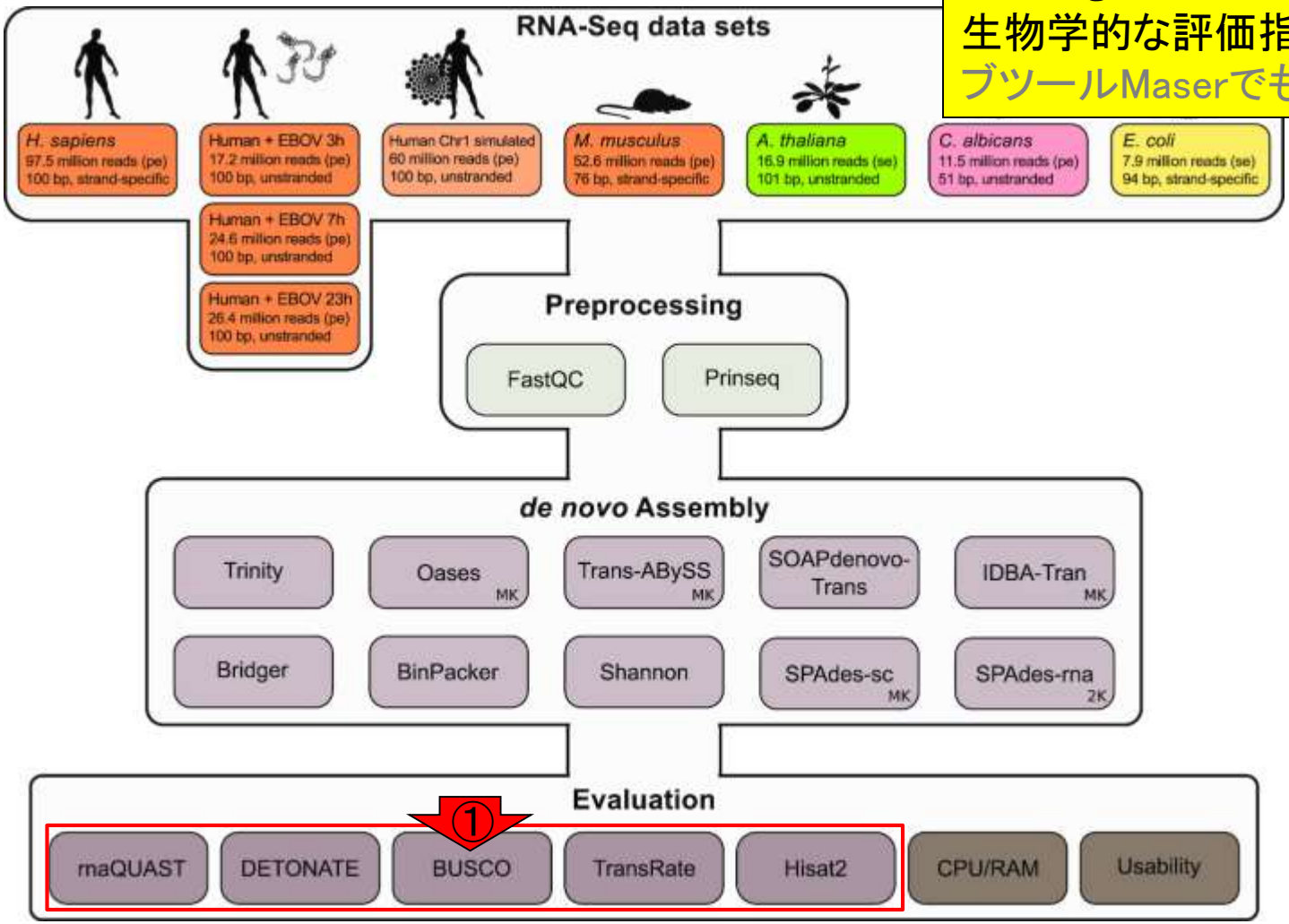
Mapping rate

We used HISAT2 v2.0.4 (RRID:SCR_015④) [41] to map the quality-controlled reads back to each assembly. The re-mapping rate can give preliminary insights into the quality of a transcriptome assembly (Fig. S4); however, further metrics are needed to assess a more complete picture of each assembler's performance.

⑤

手法比較論文の図1

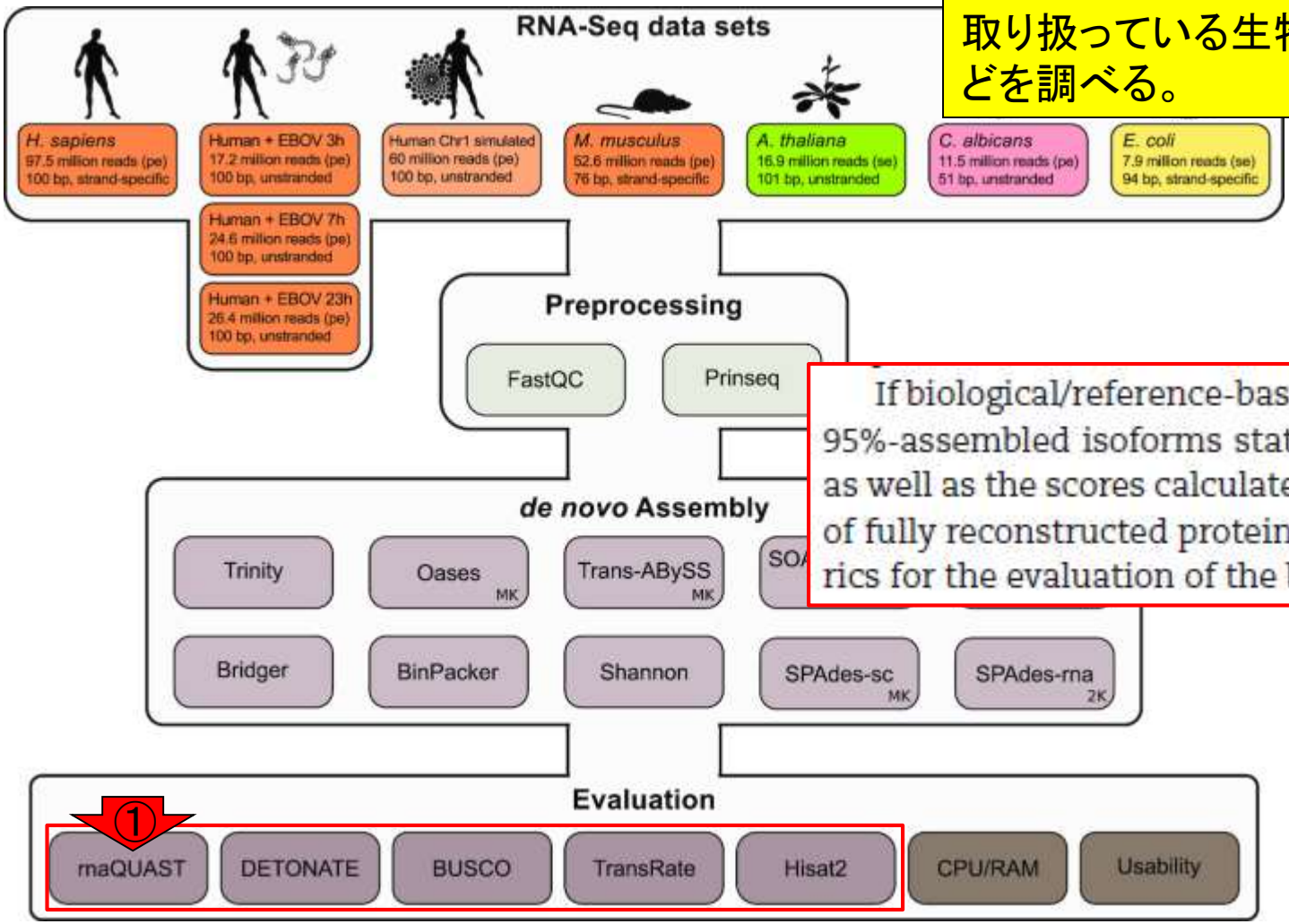
①BUSCOはゲノムアセンブリの評価指標としても用いられている。universal single-copy orthologをキチンと同定できているかどうかは、生物学的な評価指標として重要。後述するウェブツールMaserでも実行可能です。



参考

手法比較論文の図1

他の評価指標として、①rnaQUASTというものが存在することなどを知る。②原著論文を読み、自分が重視する評価指標に合致するかや、自分が取り扱っている生物種にも適用可能かどうかなどを調べる。



②

If biological/reference-based metrics should be included, 95%-assembled isoforms statistics calculated by rnaQUAST [39], as well as the scores calculated by BUSCO [43,42] and the number of fully reconstructed protein-coding transcripts, are good metrics for the evaluation of the best assembly results.

参考

手法比較論文の表4

表4が用いた評価指標。1つ1つを細かく見る必要はないが、指標としてこのようなものがあるということを大まかに記憶しておけばよいだろう。

Table 4: Selected evaluation metrics applied for each assembly and data set

No.	Tool	Selected metric	Source
1	HISAT2	Overall mapping rate	[40]
2	rnaQUAST	Transcripts ≥1,000 nt	[39]
3*		Misassemblies	
4*		Mismatches per transcript	
5*		Average alignment length	
6*		95%-assembled isoforms	
7*		Duplication ratio	
8	Trinity/Salmon	Ex90N50[a]	[10,53]
9*	Trinity/Blastx	Full-length transcripts[b]	[10,48]
10*	TransRate	Reference coverage	[42]
11		Mean ORF percentage	
12		Optimal score[c]	
13		Percentage bases uncovered[c]	
14		Number of ambiguous bases	
15	DETONATE	Nucleotide F1	[41]
16		Contig F1	
17		KC score	
18		RSEM-EVAL	
19*	BUSCO	Complete BUSCOs[d]	[43,42]
20*		Missing BUSCOs	

Metrics marked with an asterisk are biological/reference-based. All other metrics only rely on the reads used to build the assembly and/or the resulting contigs. Details can be found in the Methods. ORF: open reading frame.

[a]N50 statistic limited to the most highly expressed transcripts, which account for 90% of the total normalized expression data, calculated with the Trinity toolkit utilities.

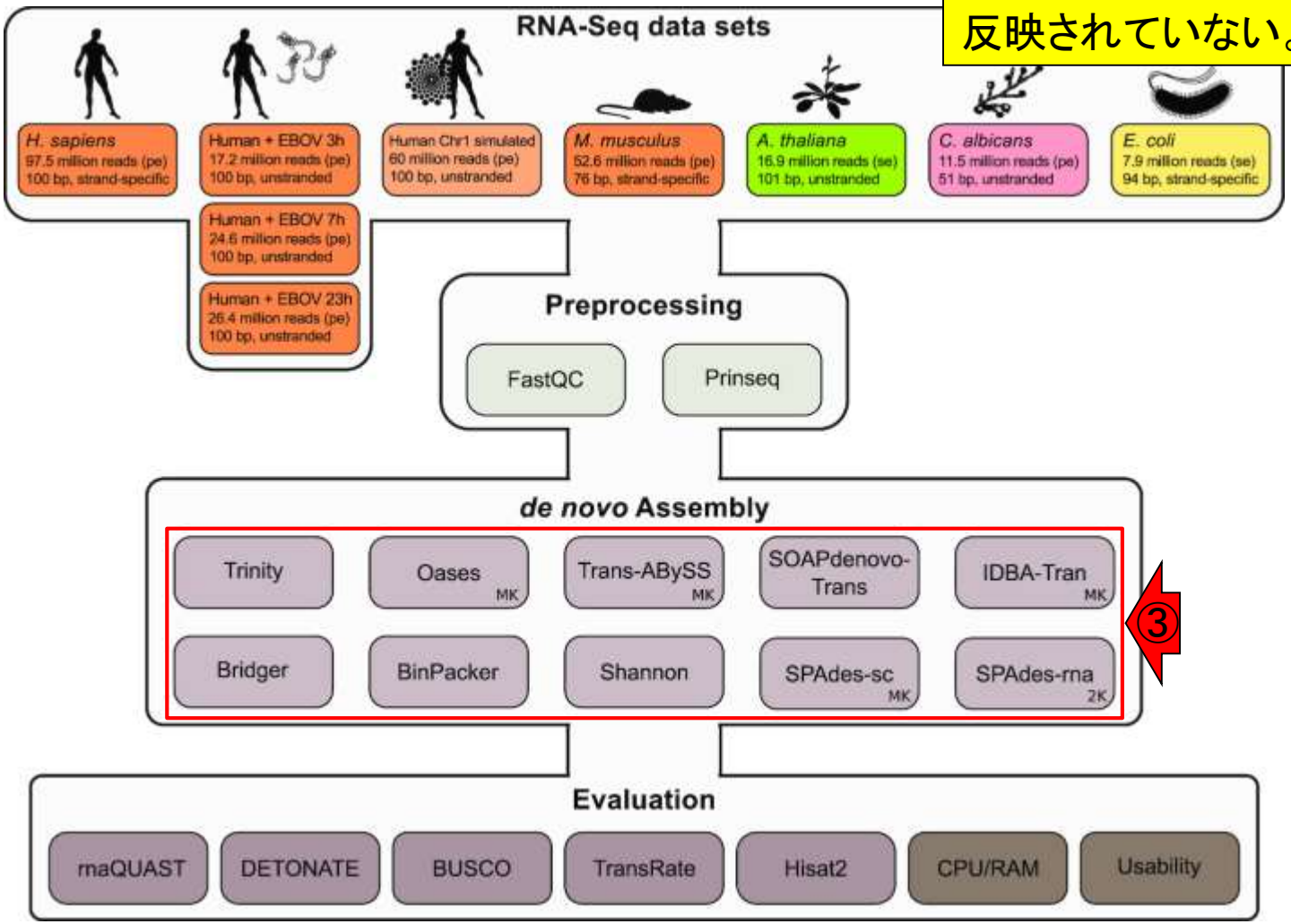
[b]Number of proteins covered by >90% by assembled transcripts.

[c]Not available for the *E. coli* and *A. thaliana* data sets because only calculated by TransRate if paired-end data are available.

[d]Sum of complete single-copy and complete duplicated BUSCOs.

手法比較論文の図1

重要事項として、この手法比較論文は2019年5月に出たものだが、①比較対象プログラムは2016年までのもの。つまり最新のプログラムは反映されていない。



最新のプログラム

例えば①2019年4月に出たTransLiGが本当に実用的なものどうかは不明だが、少なくとも試してみる価値はあるだろう。②で示されたURLからプログラムをダウンロードしてLinux環境でインストールする。

Genome Biol. 2019 Apr 23;20(1):81. doi: 10.1186/s13059-019-169

TransLiG: a de novo transcriptome assembler that uses line graph iteration.

Liu J[1], Yu T[1], Mu Z[1], Li G[2].

Author information

Abstract

We present TransLiG, a new de novo transcriptome assembler, which is able to integrate the sequence depth and pair-end information into the assembling procedure by phasing paths and iteratively constructing line graphs starting from splicing graphs. TransLiG is shown to be significantly superior to all the salient de novo assemblers in both accuracy and computing resources when tested on artificial and real RNA-seq data. TransLiG is freely available at https://sourceforge.net/projects/transcriptomeassembly/files/ .

最新のプログラム

例えば①2019年4月に出たTransLiGが本当に実用的なものどうかは不明だが、少なくとも試してみる価値はあるだろう。②で示されたURLからプログラムをダウンロードしてLinux環境でインストールする。③が当該プログラム。当然ながらLinuxの作法（makeでインストール）を知らないとどうにもならない。

① Genome Biol. 2019 Apr 23;20(1):81. doi: 10.1186/s13059-019-169

TransLiG: a de novo transcriptome as graph iteration.

Liu J[1], Yu T[1], Mu Z[1], Li G[2].

Author information

Abstract

We present TransLiG, a nev
sequence depth and pair-er
iteratively constructing line
significantly superior to all t
resources when tested on a
https://sourceforge.net/proje

Name	Modified	Size	Downloads / Week
Feature Extraction	2019-04-27		0
Baseline_Correction	2015-01-14		2
TransLiG_1.0.tar.gz ③	2018-08-12	2.4 MB	5
TransComb_v.1.0.tar.gz	2018-04-06	4.0 MB	2
Readme.txt	2016-11-26	1.4 kB	7
BinPacker_1.0.tar.gz	2016-10-19	2.3 MB	3
BinPacker_binary.tar.gz	2016-09-27	88.3 MB	4
Totals: 7 Items		97.1 MB	23

Contents

- トランスクリプトーム解析技術の原理や特徴
  - RNA-seq（Illuminaの場合）、遺伝子≠転写物
- RNA-seqデータ解析のイメージ
  - マッピング→新規転写物の同定
- 様々な解析目的
  - 転写物配列取得、手法比較論文の紹介、ウェブツール（Maser）
  - データ解析の全体像（入出力の関係や代表的なツール）
- アノテーションファイルの読み込みと課題1
  - Rで転写物配列取得のイントロ
- Rで転写物配列取得と課題2
  - アノテーションファイルとゲノム情報ファイルから
- 公共データベース
  - NGS全体（NCBI SRA, EMBL-EBI ENA, DDBJ SRA）
  - DRAの概要、クオリティスコアなど

ウェブツールMaser

遺伝研（静岡県三島市）で運用されているクラウド解析環境である①Maserは、Linux-free（データをアップロードしてボタンをポチポチ押していけば解析できるという意味）のNGSデータ解析手段です。

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed ▾ Advanced Help

Format: Abstract ▾ Send to ▾

Database (Oxford). 2018 Jan 1;2018. doi: 10.1093/database/bay027.

Maser: one-stop platform for NGS big data from analysis to visualization.

Kinjo S[1], Monma N[2], Misu S[3], Kitamura N[1], Imoto J[1], Yoshitake K[4], Gojobori T[5], Ikeo K[1,6].

⊕ Author information

Abstract

http://cell-innovation.nig.ac.jp/maser/.

PMID: 29688385 PMCID: PMC5905357 DOI: 10.1093/database/bay027

[Indexed for MEDLINE] Free PMC Article

Images from this publication. See all images (3) Free text



Full text links

① OXFORD ACADEMIC

PMC FREE Full text

Save items

Add to Favorites

Similar articles

Epigenomic annotation of genetic variants using [Nat Biotechnol. 2015]

Epiviz: interactive visual analytics for functional genon [Nat Methods. 2014]

D3GB: An Interactive Genome Browser for R, [J Comput Biol. 2017]

Review Cytogenetic Resources and Information. [Methods Mol Biol. 2017]

Review Implementing WebGL and HTML5 in ! [Trends Biotechnol. 2017]

ウェブツールMaser

遺伝研（静岡県三島市）で運用されているクラウド解析環境である①Maserは、Linux-free（データをアップロードしてボタンをポチポチ押していけば解析できるという意味）のNGSデータ解析手段です。②なぜかAbstractがMaserのURL情報のみになってしまっていますが、③Full textのリンク先には正しいアブストラクトがあります。

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed ▾ Advanced

Format: Abstract ▾

Database (Oxford). 2018 Jan 1;2018. doi: 10.1093/database/bay027.

Maser: one-stop platform for NGS big data from analysis to visualization.

Kinjo S[1], Monma N[2], Misu S[3], Kitamura N[1], Imoto J[1], Yoshitake K[4], Gojobori T[5], Ikeo K[1,6].

⊕ Author information

Abstract

http://cell-innovation.nig.ac.jp/maser/.

PMID: 29688385 PMCID: PMC5905357 DOI: 10.1093/database/bay027

[Indexed for MEDLINE] Free PMC Article

Images from this publication. See all images (3) Free text



① ② ③

OXFORD ACADEMIC

PMC FREE Full text

Save items

Add to Favorites

Similar articles

Epigenomic annotation of genetic variants using [Nat Biotechnol. 2015]

Epiviz: interactive visual analytics for functional genon [Nat Methods. 2014]

D3GB: An Interactive Genome Browser for R, [J Comput Biol. 2017]

Review Cytogenetic Resources and Information. [Methods Mol Biol. 2017]

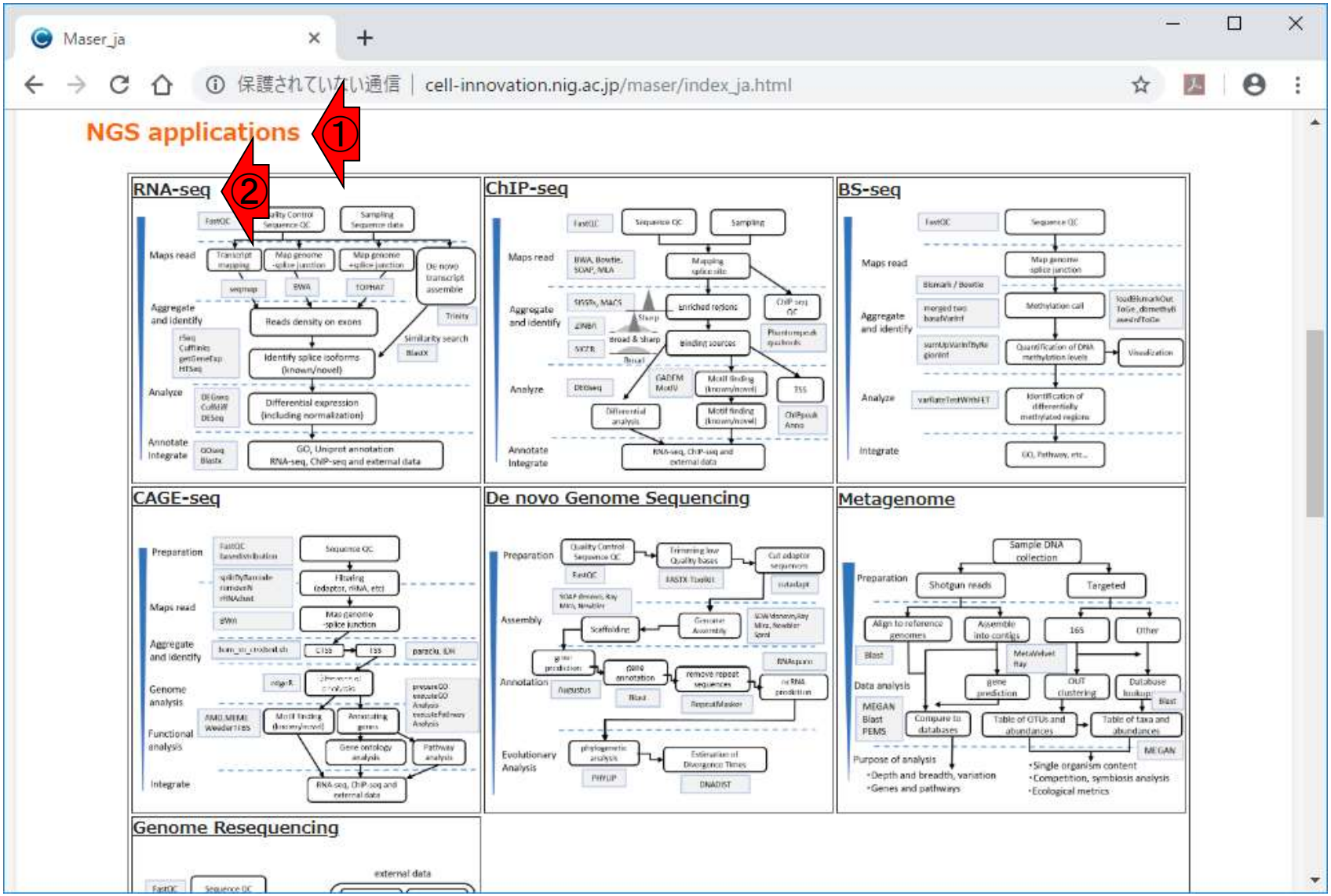
Review Implementing WebGL and HTML5 in ! [Trends Biotechnol. 2017]

ウェブツールMaser

Maserのトップページ（日本語版）。①で示されているように、様々な解析が可能。②ページ下部に移動。

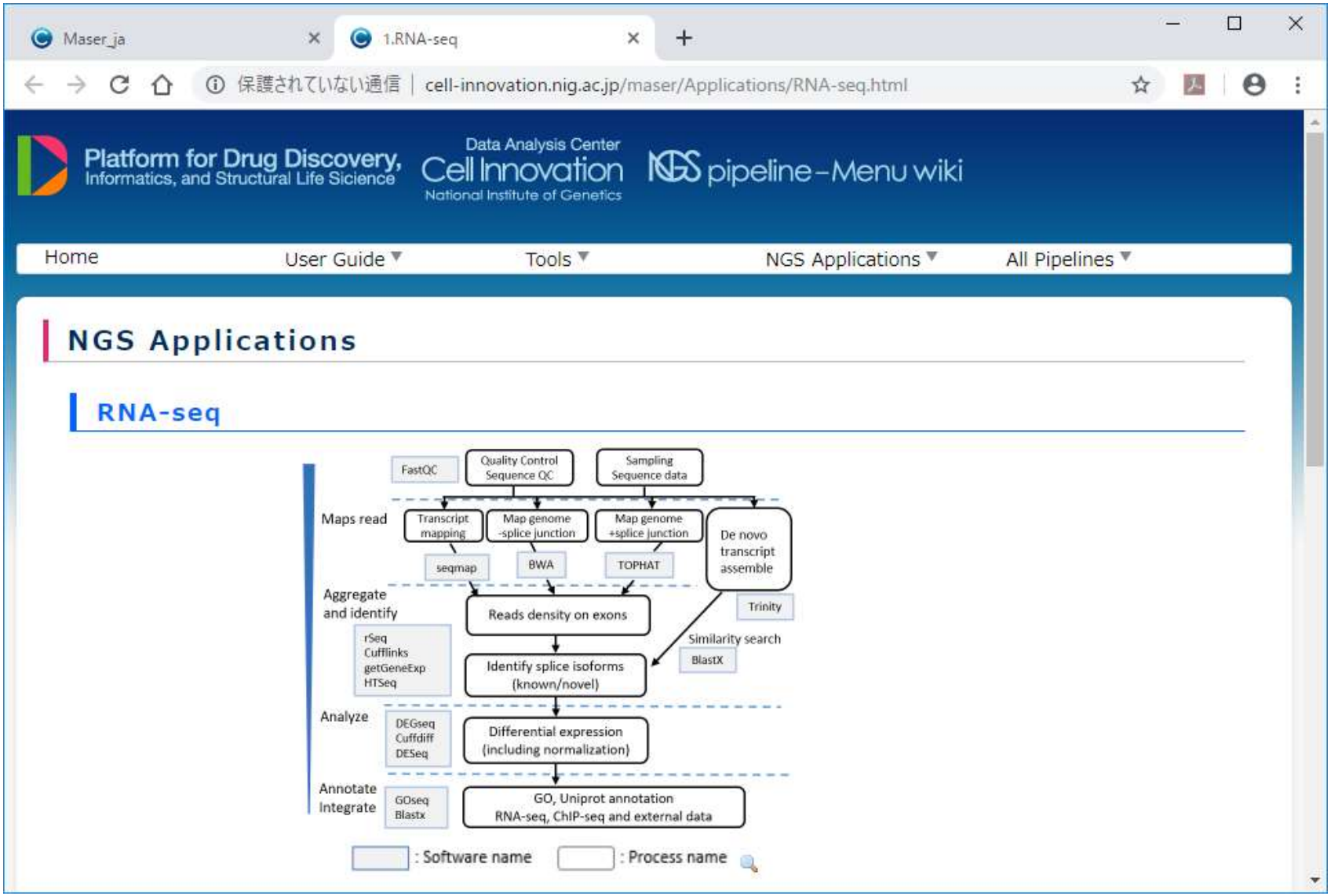
ウェブツールMaser

①このあたりまで移動。例えば②RNA-seqをクリック。



ウェブツールMaser

①このあたりまで移動。例えば②RNA-seqをクリック。こんな感じになります。



ウェブツールMaser

①このあたりまで移動。例えば②RNA-seqをクリック。こんな感じになります。③様々なプログラムを利用できそうなことがわかります。④ページ下部に移動。

Maser_ja | 1.RNA-seq

保護されていない通信 | cell-innovation.nig.ac.jp/maser/Applications/RNA-seq.html

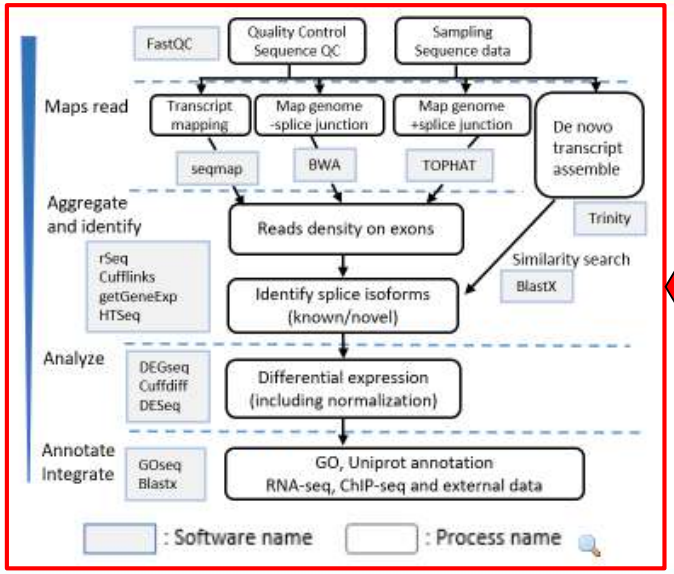
Platform for Drug Discovery, Informatics, and Structural Life Science
Data Analysis Center Cell Innovation National Institute of Genetics
NGS pipeline-Menu wiki

Home | User Guide | Tools | NGS Applications | All Pipelines

④

NGS Applications

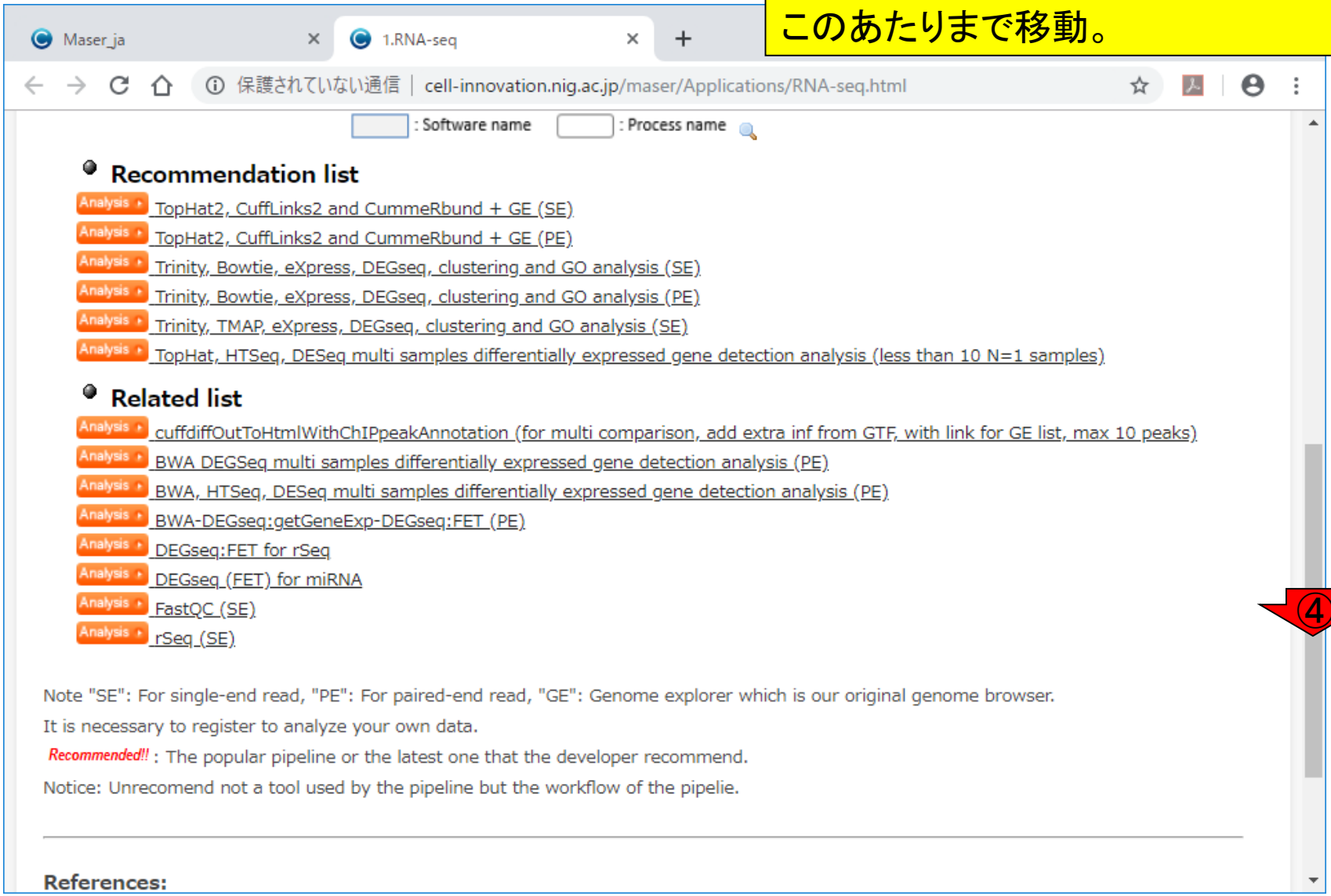
RNA-seq



③

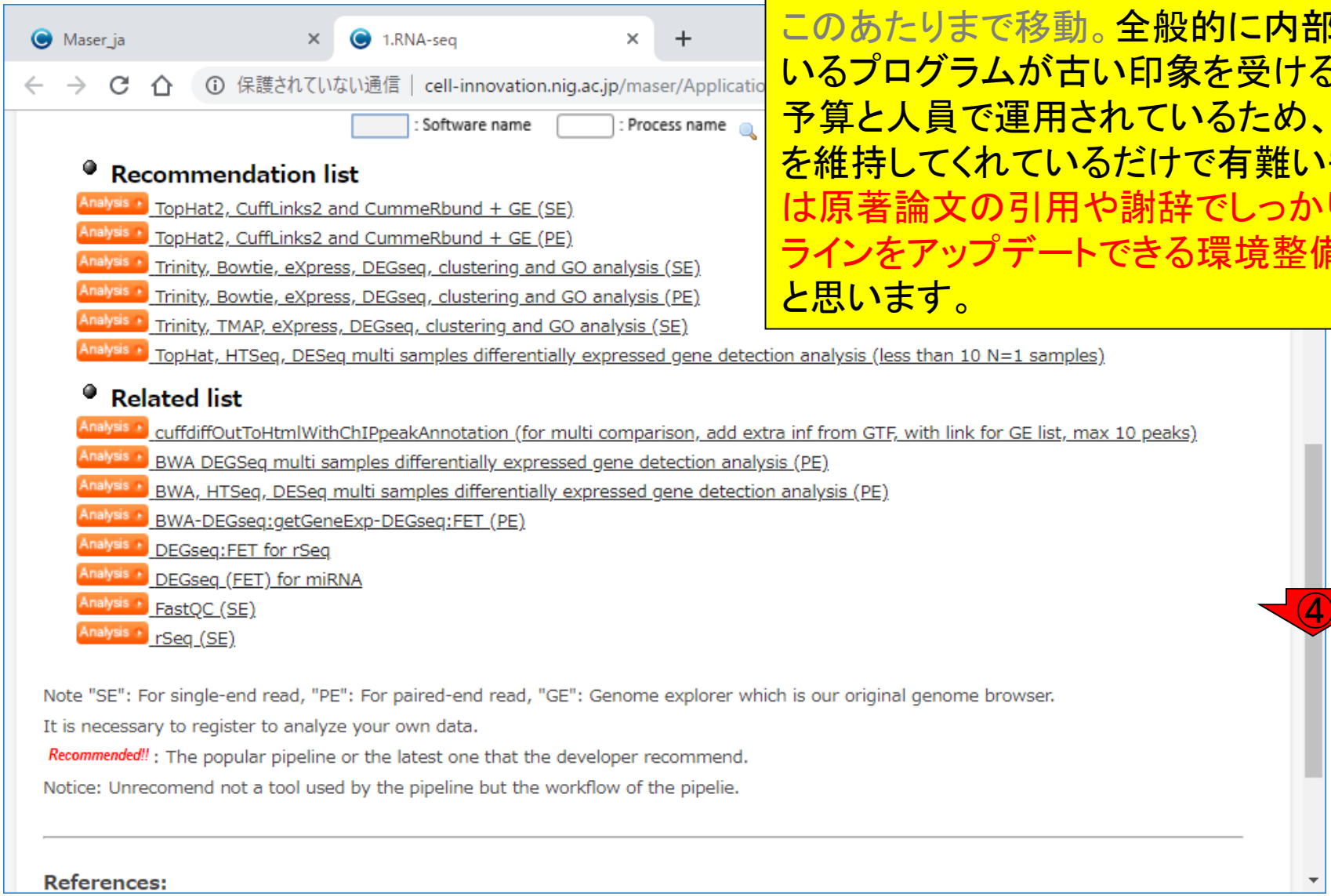
ウェブツールMaser

①このあたりまで移動。例えば②RNA-seqをクリック。こんな感じになります。③様々なプログラムを利用できそうなことがわかります。④ページ下部に移動。
このあたりまで移動。



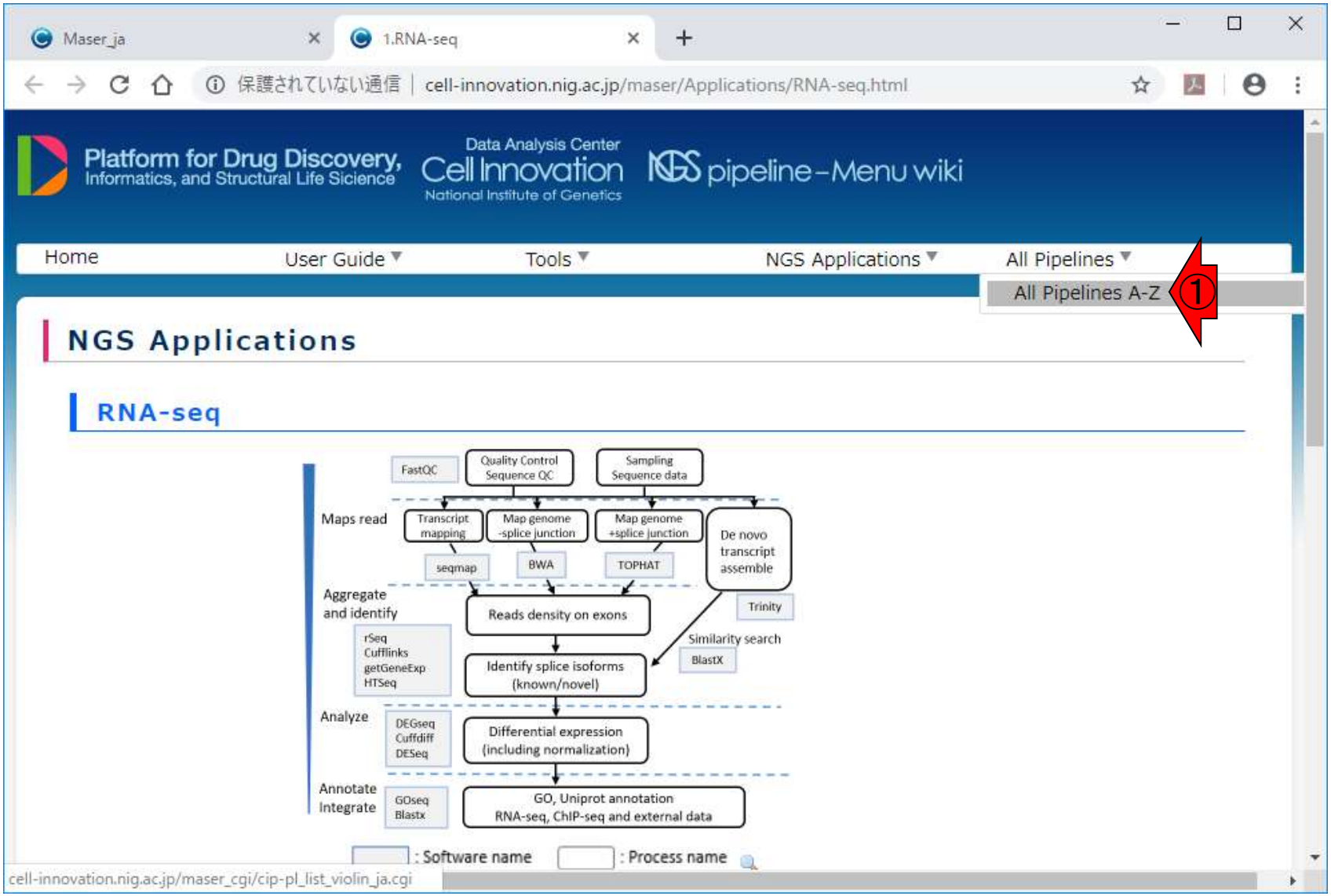
ウェブツールMaser

①このあたりまで移動。例えば②RNA-seqをクリック。こんな感じになります。③様々なプログラムを利用できそうなことがわかります。④ページ下部に移動。このあたりまで移動。全般的に内部的に利用されているプログラムが古い印象を受ける。しかし、少ない予算と人員で運用されているため、利用できる状況を維持してくれているだけで有難いものです。ユーザは原著論文の引用や謝辞でしっかり応援し、パイプラインをアップデートできる環境整備に尽力すべきだと思います。



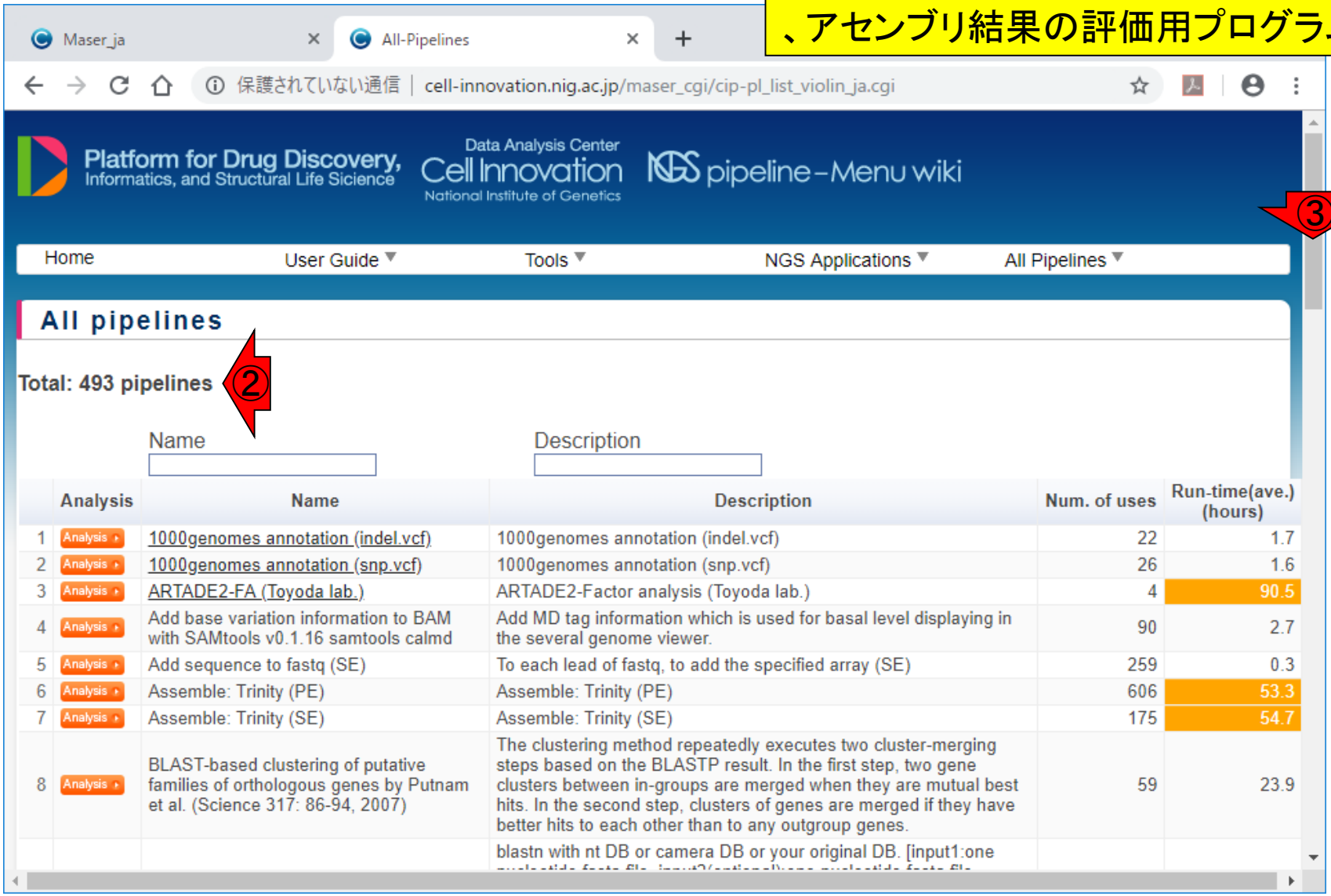
ウェブツールMaser

①All Pipelinesでどのような解析パイプラインを利用可能か知ることができます。



ウェブツールMaser

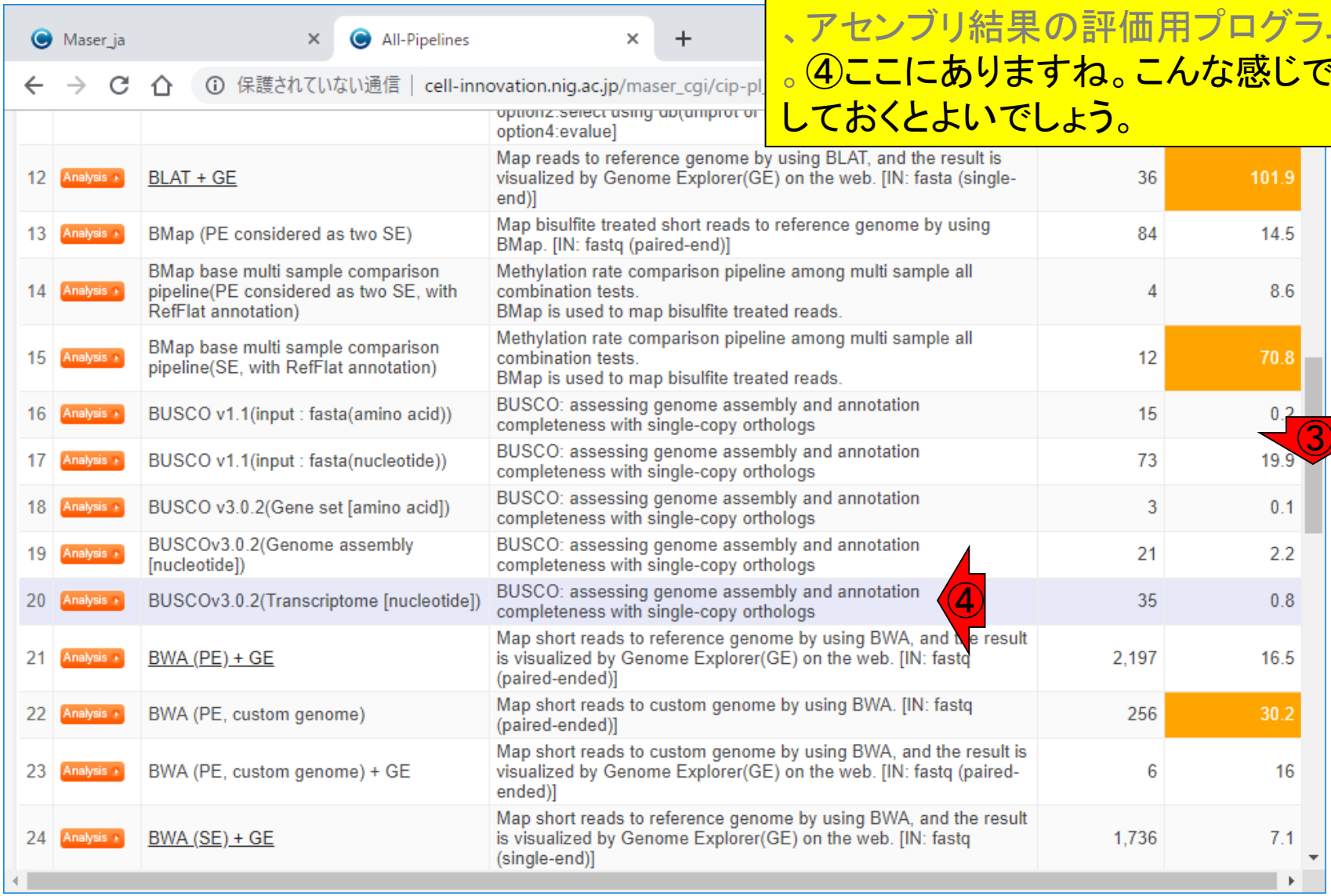
①All Pipelinesでどのような解析パイプラインを利用可能か知ることができます。こんな感じになります。②トータルで493個あります。③ページ下部に移動し、アセンブリ結果の評価用プログラムBUSCOを探す



	Analysis	Name	Description	Num. of uses	Run-time(ave.) (hours)
1	Analysis	1000genomes annotation (indel.vcf)	1000genomes annotation (indel.vcf)	22	1.7
2	Analysis	1000genomes annotation (snp.vcf)	1000genomes annotation (snp.vcf)	26	1.6
3	Analysis	ARTADE2-FA (Toyoda lab.)	ARTADE2-Factor analysis (Toyoda lab.)	4	90.5
4	Analysis	Add base variation information to BAM with SAMtools v0.1.16 samtools calmd	Add MD tag information which is used for basal level displaying in the several genome viewer.	90	2.7
5	Analysis	Add sequence to fastq (SE)	To each lead of fastq, to add the specified array (SE)	259	0.3
6	Analysis	Assemble: Trinity (PE)	Assemble: Trinity (PE)	606	53.3
7	Analysis	Assemble: Trinity (SE)	Assemble: Trinity (SE)	175	54.7
8	Analysis	BLAST-based clustering of putative families of orthologous genes by Putnam et al. (Science 317: 86-94, 2007)	The clustering method repeatedly executes two cluster-merging steps based on the BLASTP result. In the first step, two gene clusters between in-groups are merged when they are mutual best hits. In the second step, clusters of genes are merged if they have better hits to each other than to any outgroup genes.	59	23.9
			blastn with nt DB or camera DB or your original DB. [input1:one		

ウェブツールMaser

①All Pipelinesでどのような解析パイプラインを利用可能か知ることができます。こんな感じになります。②トータルで493個あります。③ページ下部に移動し、アセンブリ結果の評価用プログラムBUSCOを探す。④ここにありますね。こんな感じでできることを概観しておくとよいでしょう。



12	Analysis	BLAT + GE	Map reads to reference genome by using BLAT, and the result is visualized by Genome Explorer(GE) on the web. [IN: fasta (single-end)]	36	101.9
13	Analysis	BMap (PE considered as two SE)	Map bisulfite treated short reads to reference genome by using BMap. [IN: fastq (paired-end)]	84	14.5
14	Analysis	BMap base multi sample comparison pipeline(PE considered as two SE, with RefFlat annotation)	Methylation rate comparison pipeline among multi sample all combination tests. BMap is used to map bisulfite treated reads.	4	8.6
15	Analysis	BMap base multi sample comparison pipeline(SE, with RefFlat annotation)	Methylation rate comparison pipeline among multi sample all combination tests. BMap is used to map bisulfite treated reads.	12	70.8
16	Analysis	BUSCO v1.1(input : fasta(amino acid))	BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs	15	0.2
17	Analysis	BUSCO v1.1(input : fasta(nucleotide))	BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs	73	19.9
18	Analysis	BUSCO v3.0.2(Gene set [amino acid])	BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs	3	0.1
19	Analysis	BUSCOv3.0.2(Genome assembly [nucleotide])	BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs	21	2.2
20	Analysis	BUSCOv3.0.2(Transcriptome [nucleotide])	BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs	35	0.8
21	Analysis	BWA (PE) + GE	Map short reads to reference genome by using BWA, and the result is visualized by Genome Explorer(GE) on the web. [IN: fastq (paired-ended)]	2,197	16.5
22	Analysis	BWA (PE, custom genome)	Map short reads to custom genome by using BWA. [IN: fastq (paired-ended)]	256	30.2
23	Analysis	BWA (PE, custom genome) + GE	Map short reads to custom genome by using BWA, and the result is visualized by Genome Explorer(GE) on the web. [IN: fastq (paired-ended)]	6	16
24	Analysis	BWA (SE) + GE	Map short reads to reference genome by using BWA, and the result is visualized by Genome Explorer(GE) on the web. [IN: fastq (single-end)]	1,736	7.1

Contents

- トランスクリプトーム解析技術の原理や特徴
  - RNA-seq（Illuminaの場合）、遺伝子≠転写物
- RNA-seqデータ解析のイメージ
  - マッピング→新規転写物の同定
- 様々な解析目的
  - 転写物配列取得、手法比較論文の紹介、ウェブツール（Maser）
  - データ解析の全体像（入出力の関係や代表的なツール）
- アノテーションファイルの読み込みと課題1
  - Rで転写物配列取得のイントロ
- Rで転写物配列取得と課題2
  - アノテーションファイルとゲノム情報ファイルから
- 公共データベース
  - NGS全体（NCBI SRA, EMBL-EBI ENA, DDBJ SRA）
  - DRAの概要、クオリティスコアなど

とあるガイドライン系論文

RNA-seqデータを入力として実行する包括的な解析パイプラインRNACocktailを提唱した論文

Nat Commun. 2017 Jul 5;8(1):59. doi: 10.1038/s41467-017-00050-4.

Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis.

Sahraeian SME[1], Mohiyuddin M[1], Sebra R[2], Tilgner H[3], Afshar PT[4], Au KF[5], Bani Asadi N[1], Gerstein MB[6], Wong WH[7], Snyder MP[3], Schadt E[2], Lam HYK[8].

Author information

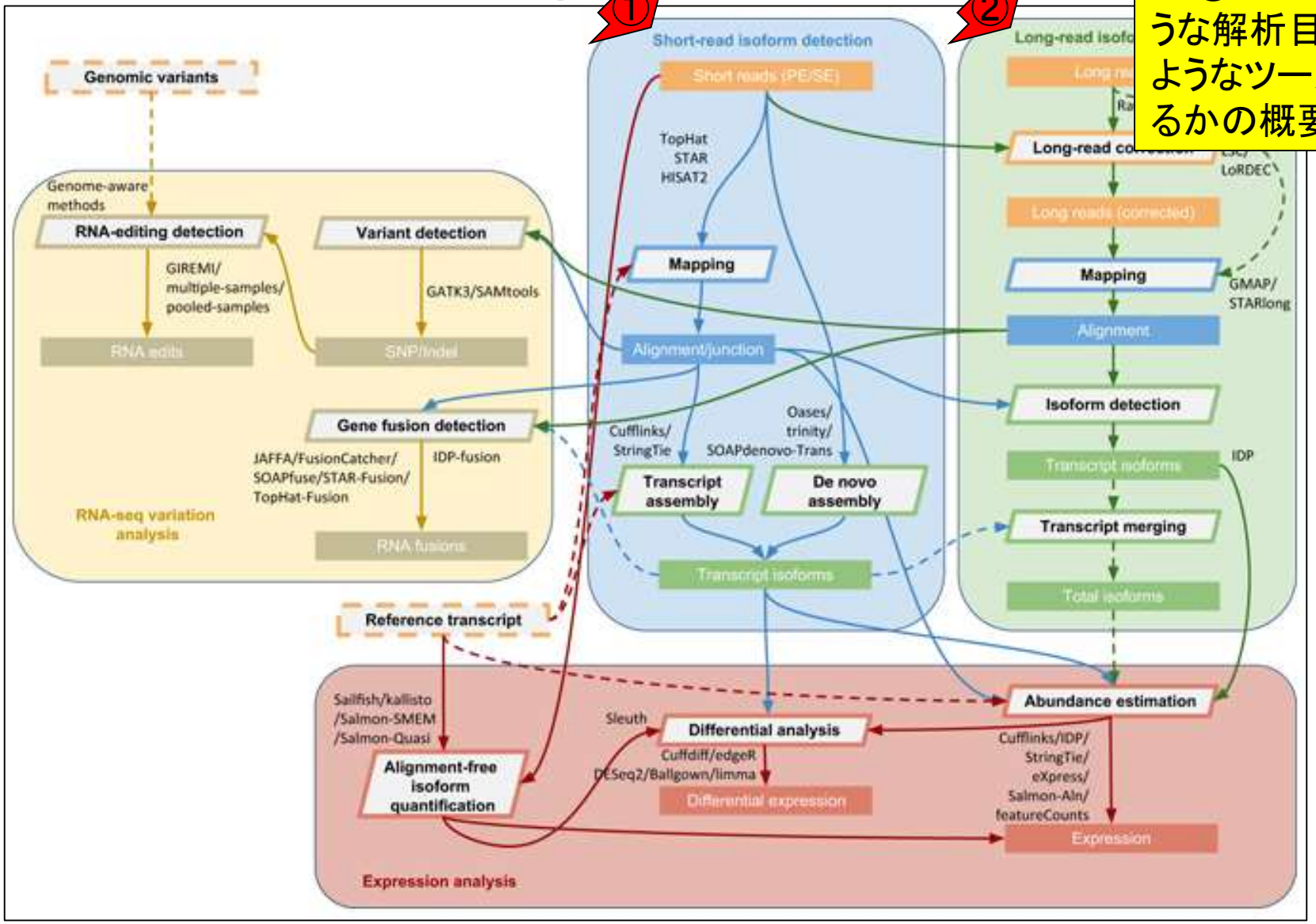
Abstract

RNA-sequencing (RNA-seq) is an essential technique for transcriptome studies, hundreds of analysis tools have been developed since it was debuted. Although recent efforts have attempted to assess the latest available tools, they have not evaluated the analysis workflows comprehensively to unleash the power within RNA-seq. Here we conduct an extensive study analysing a broad spectrum of RNA-seq workflows. Surpassing the expression analysis scope, our work also includes assessment of RNA variant-calling, RNA editing and RNA fusion detection techniques. Specifically, we examine both short- and long-read RNA-seq technologies, 39 analysis tools resulting in ~120 combinations, and ~490 analyses involving 15 samples with a variety of germline, cancer and stem cell data sets. We report the performance and propose a comprehensive RNA-seq analysis protocol, named RNACocktail, along with a computational pipeline achieving high accuracy. Validation on different samples reveals that our proposed protocol could help researchers extract more biologically relevant predictions by broad analysis of the transcriptome.RNA-seq is widely used for transcriptome analysis. Here, the authors analyse a wide spectrum of RNA-seq workflows and present a comprehensive analysis protocol named RNACocktail as well as a computational pipeline leveraging the widely used tools for accurate RNA-seq analysis.

Sahraeian et al., Nat Commun., 8(1): 59, 2017

RNACocktail論文の図1

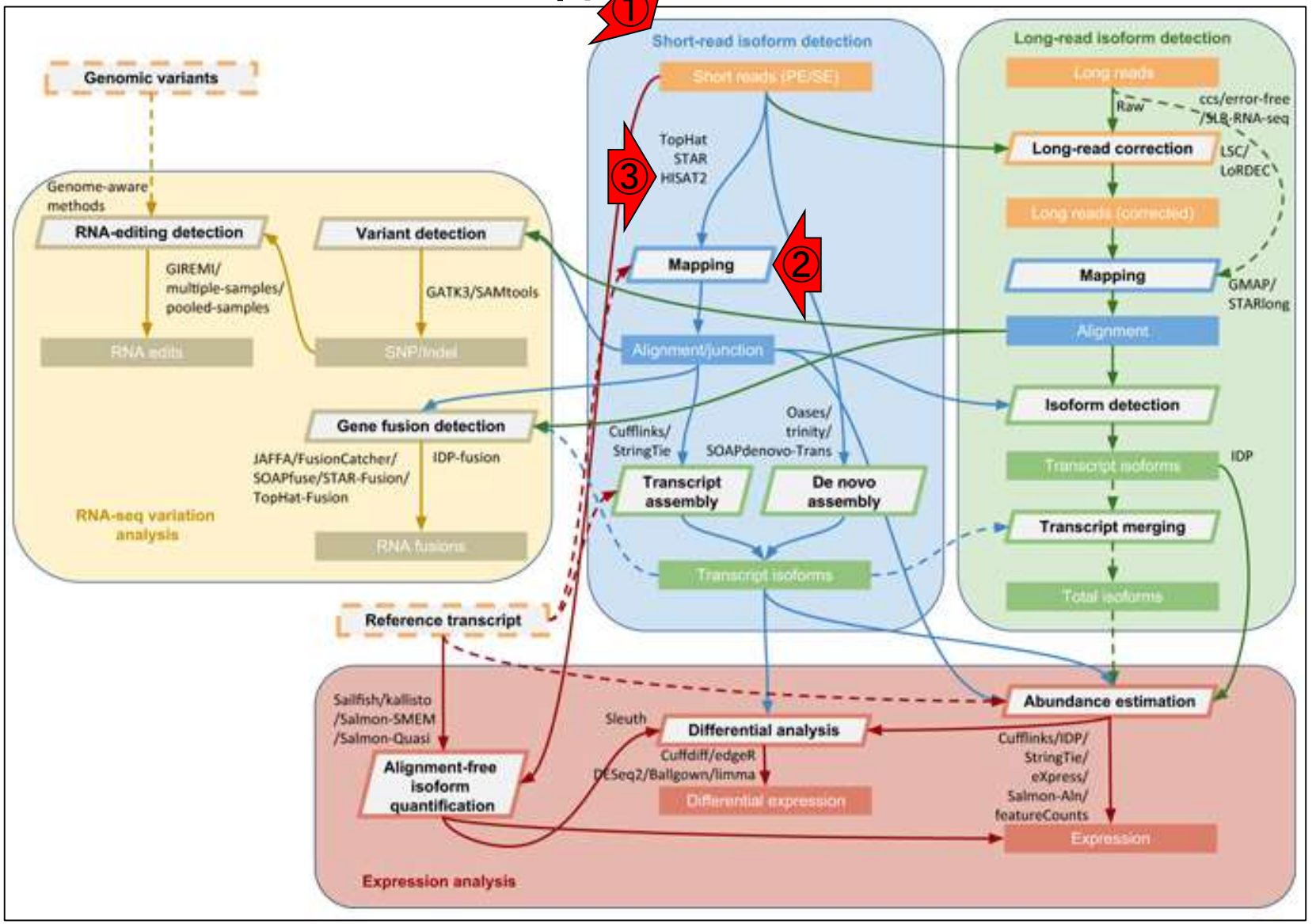
論文の図1を見ることで、どのようなデータ（①ショートリード or ②ロングリード）で、どのような解析目的の場合に、どのようなツールが用いられているかの概要がわかる



Sahraeian et al., Nat Commun., **8**(1): 59, 2017

RNACocktail論文の図1

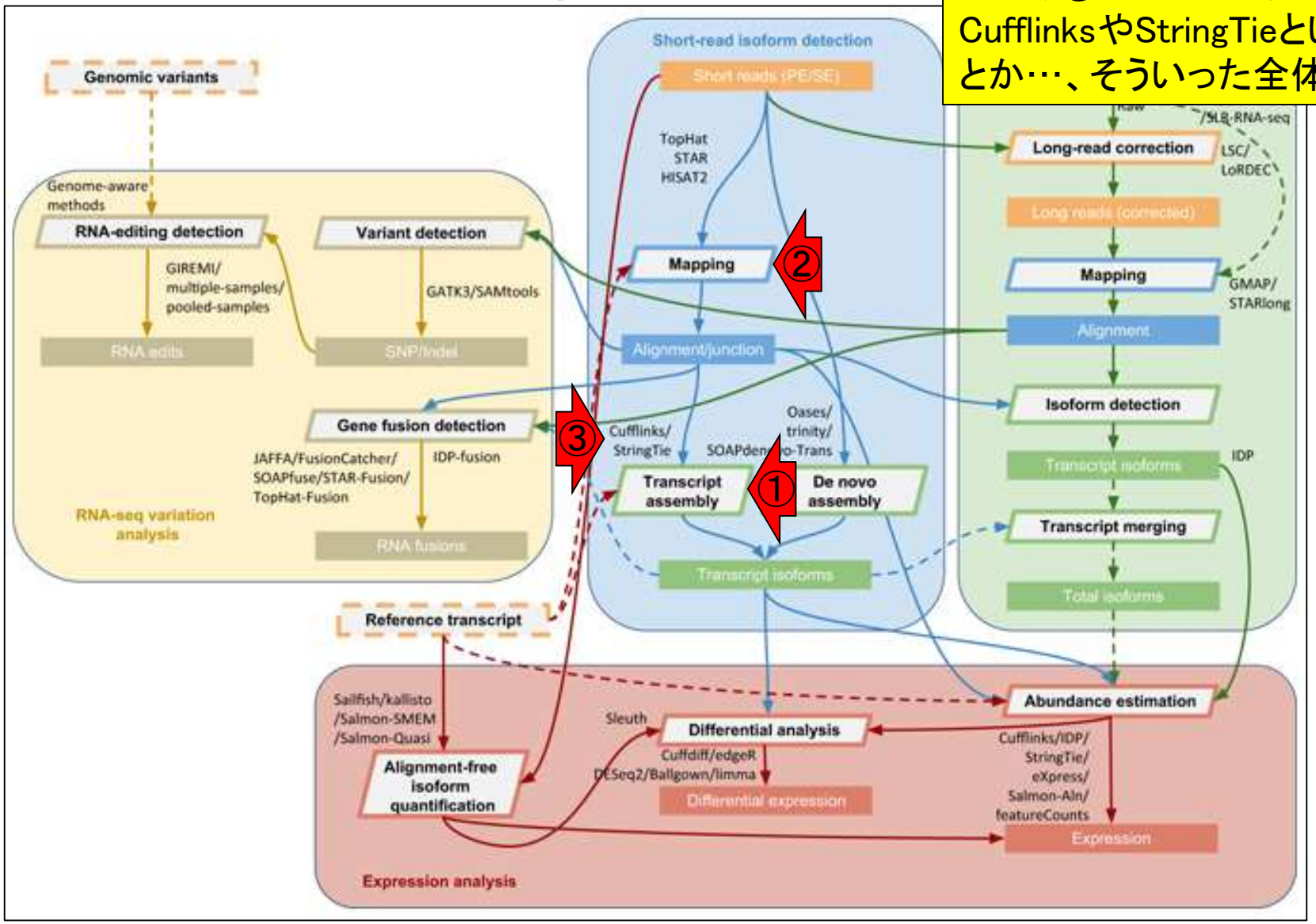
例えば、①ショートリードの②マッピングの場合は、③HISAT2などが使われるとか…



Sahraeian et al., Nat Commun., 8(1): 59, 2017

RNACocktail論文の図1

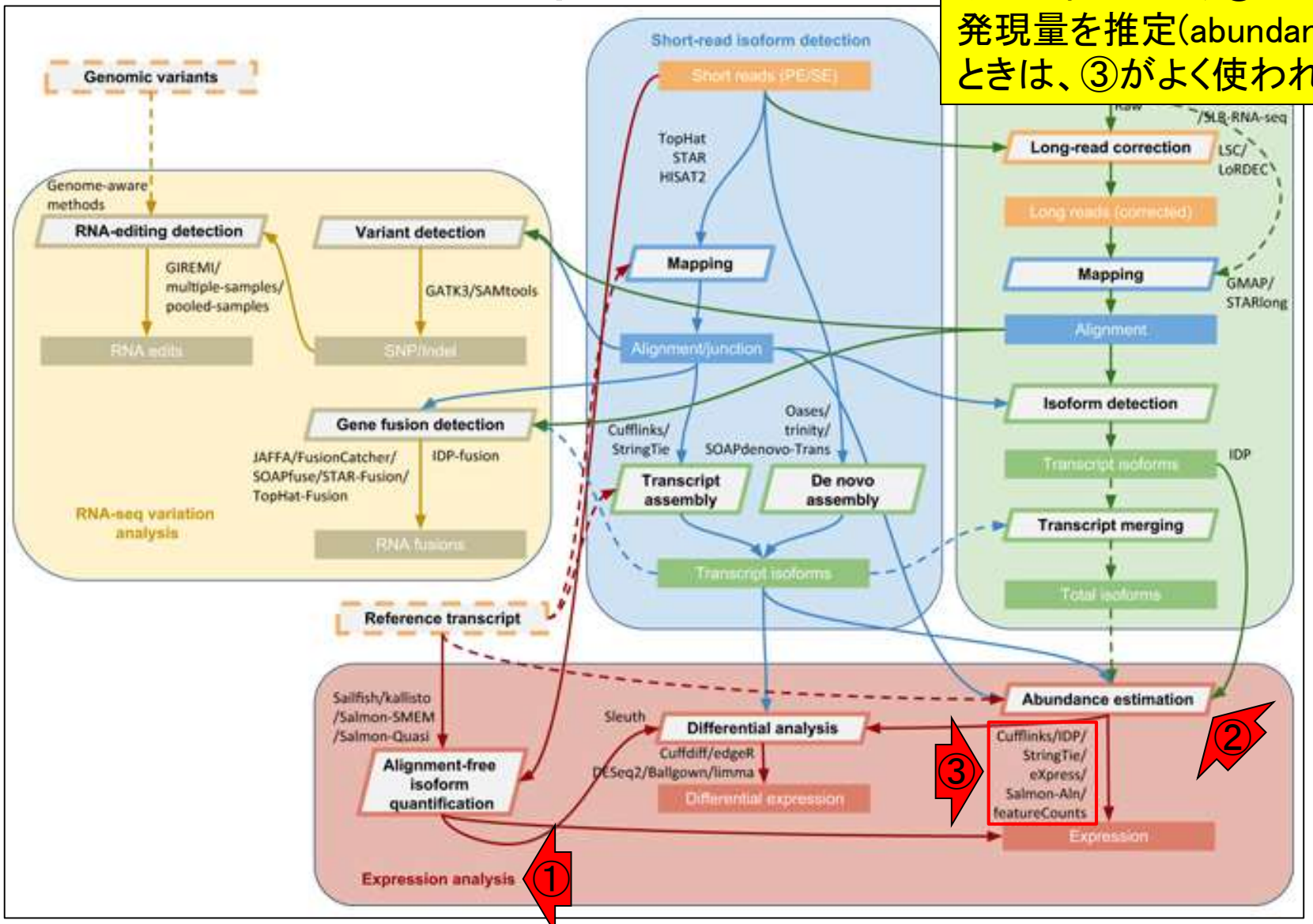
(ゲノム配列**既知**の場合で)①トランスクリプトーム配列取得(遺伝子構造推定)の場合は、②マッピング結果を入力として、③CufflinksやStringTieというツールを用いるとか…、そういった全体像がわかります



Sahraeian et al., Nat Commun., **8**(1): 59, 2017

RNACocktail論文の図1

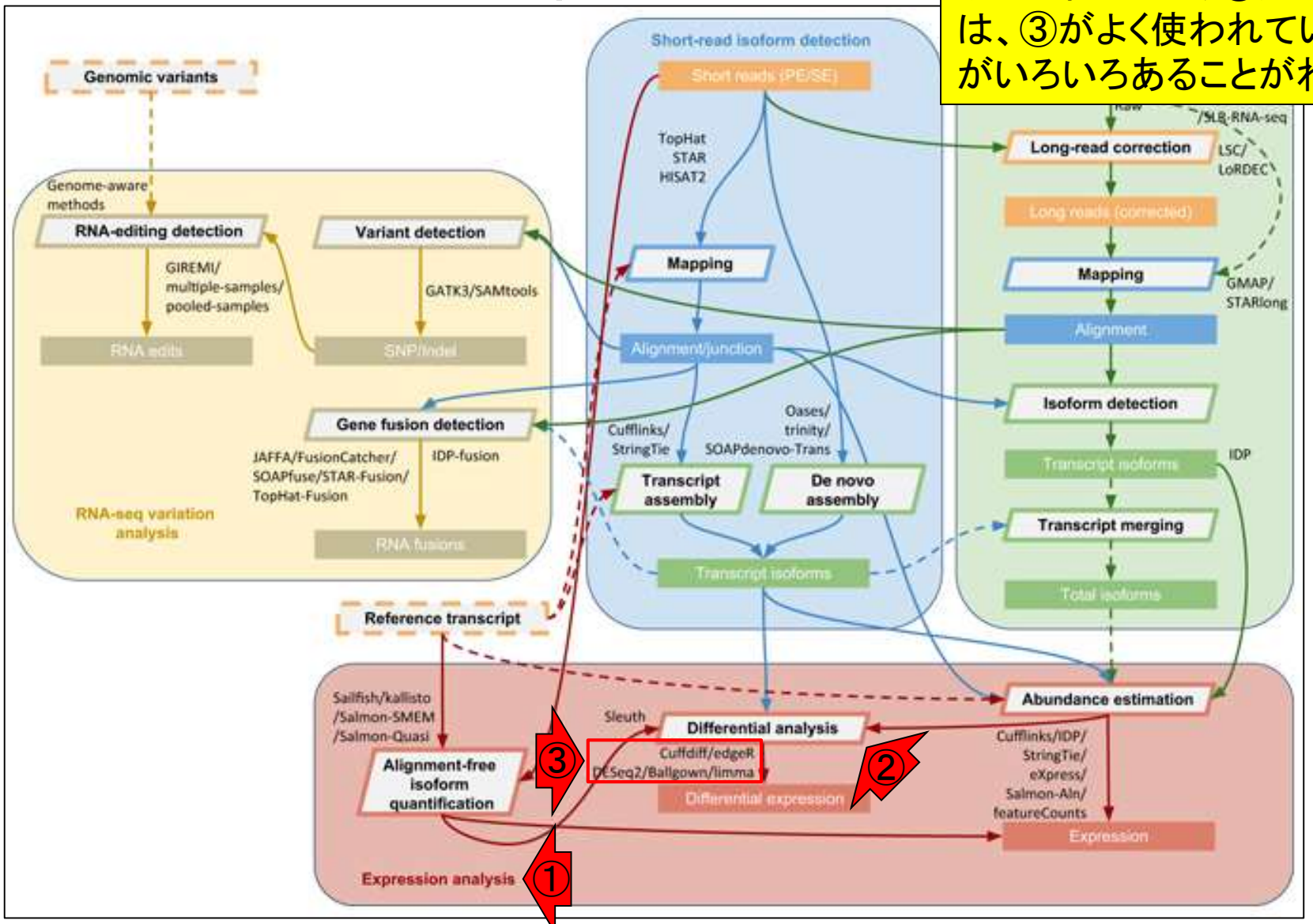
後半の第3-4回では、①発現解析のトピックを中心に教えますが、一口に①発現解析とは言っても、②遺伝子/転写物ごとの発現量を推定(abundance estimation)するときは、③がよく使われ…



Sahraeian et al., Nat Commun., 8(1): 59, 2017

RNACocktail論文の図1

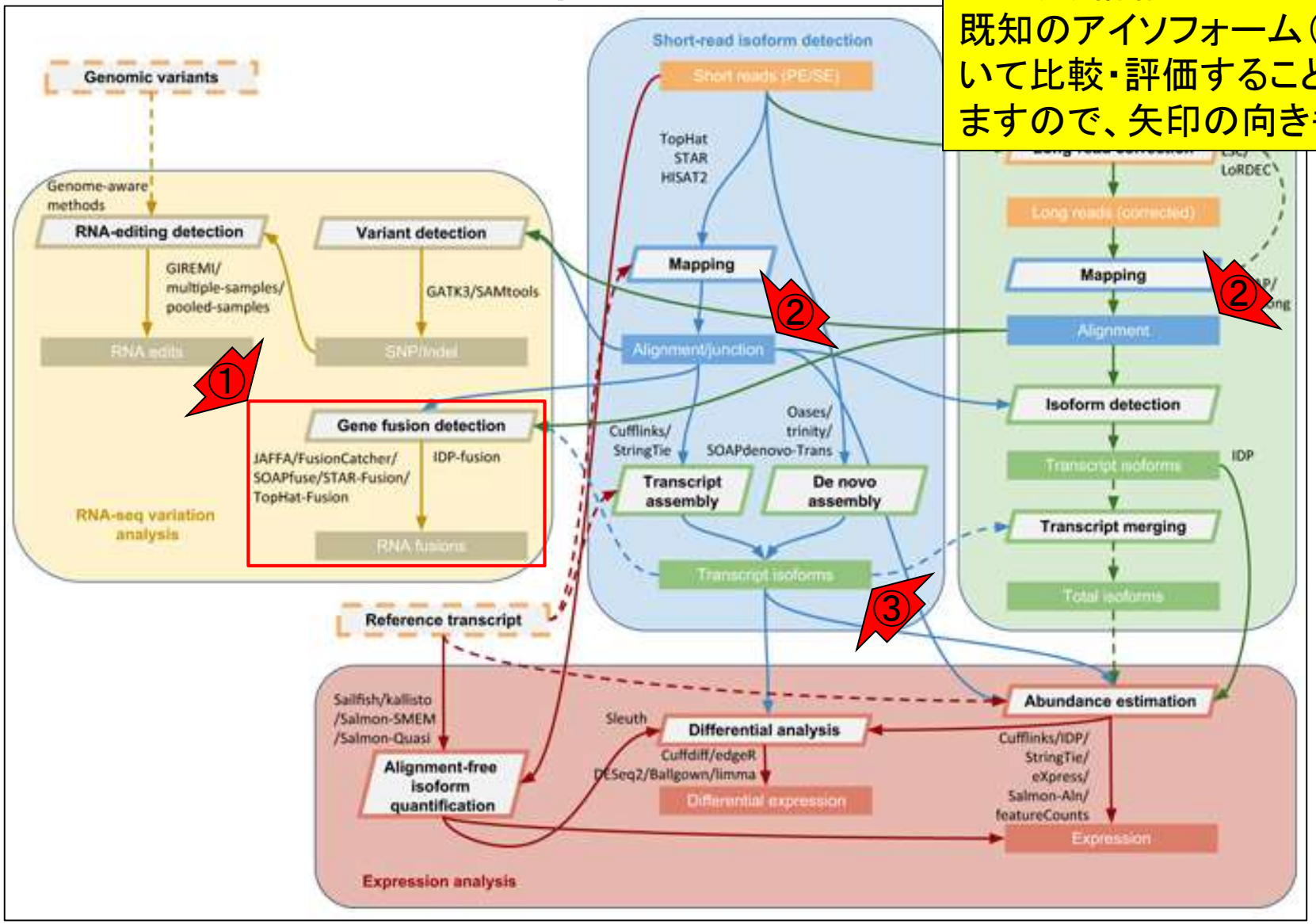
後半の第3-4回では、①発現解析のトピックを中心に教えますが、一口に①発現解析とは言っても、②発現変動解析の場合は、③がよく使われているなど、場合分けがいろいろあることがわかります



Sahraeian et al., Nat Commun., **8**(1): 59, 2017

RNACocktail論文の図1

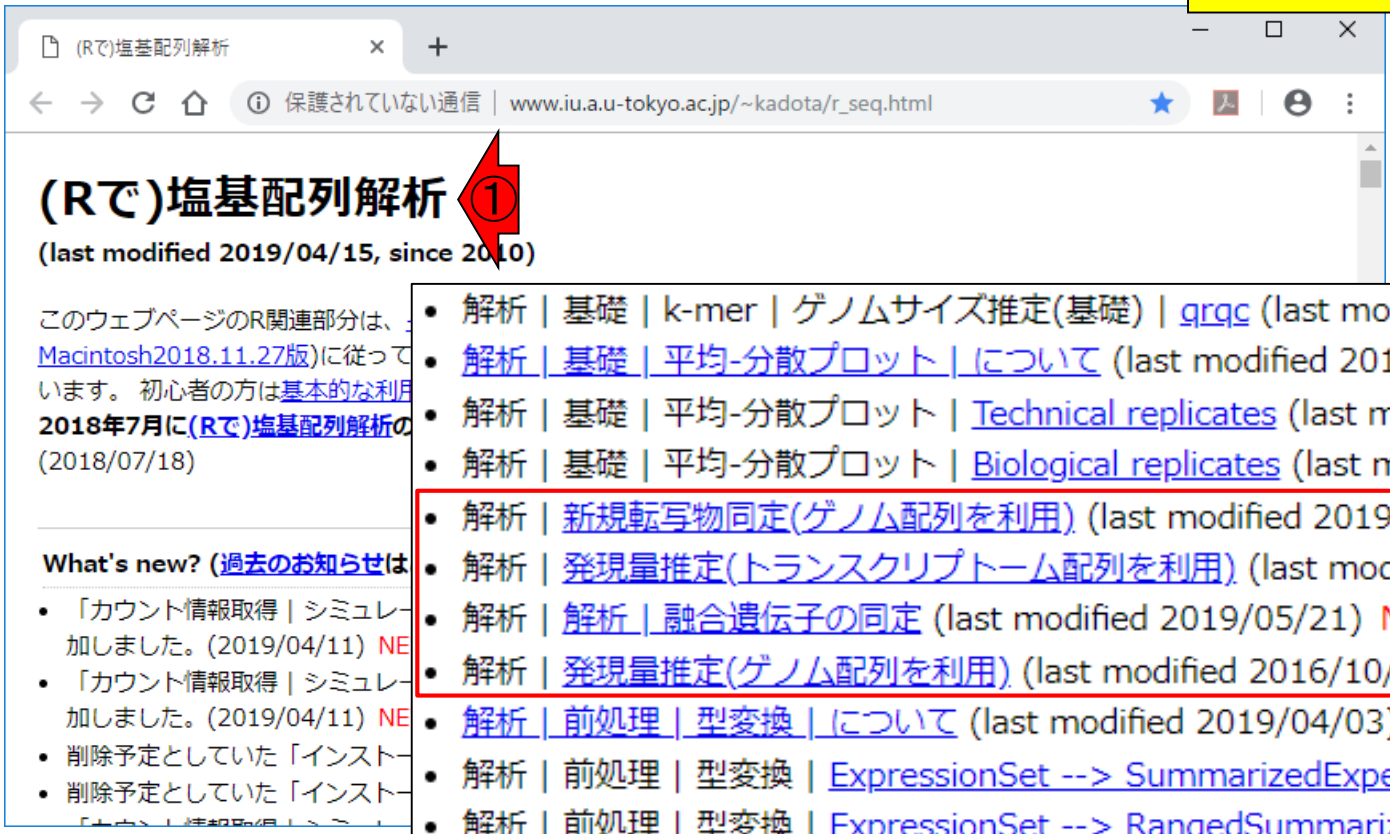
①融合遺伝子検出の場合は、②マッピング（or アラインメント）結果を入力として用います。情報として利用可能な場合は、③既知のアイソフォーム（isoform）情報を用いて比較・評価することもあろうかと思いますので、矢印の向きも妥当ですね



Sahraeian et al., Nat Commun., 8(1): 59, 2017

参考

①のページ上では、②のところに遺伝子構造推定、発現量推定、融合遺伝子検出の項目があります。



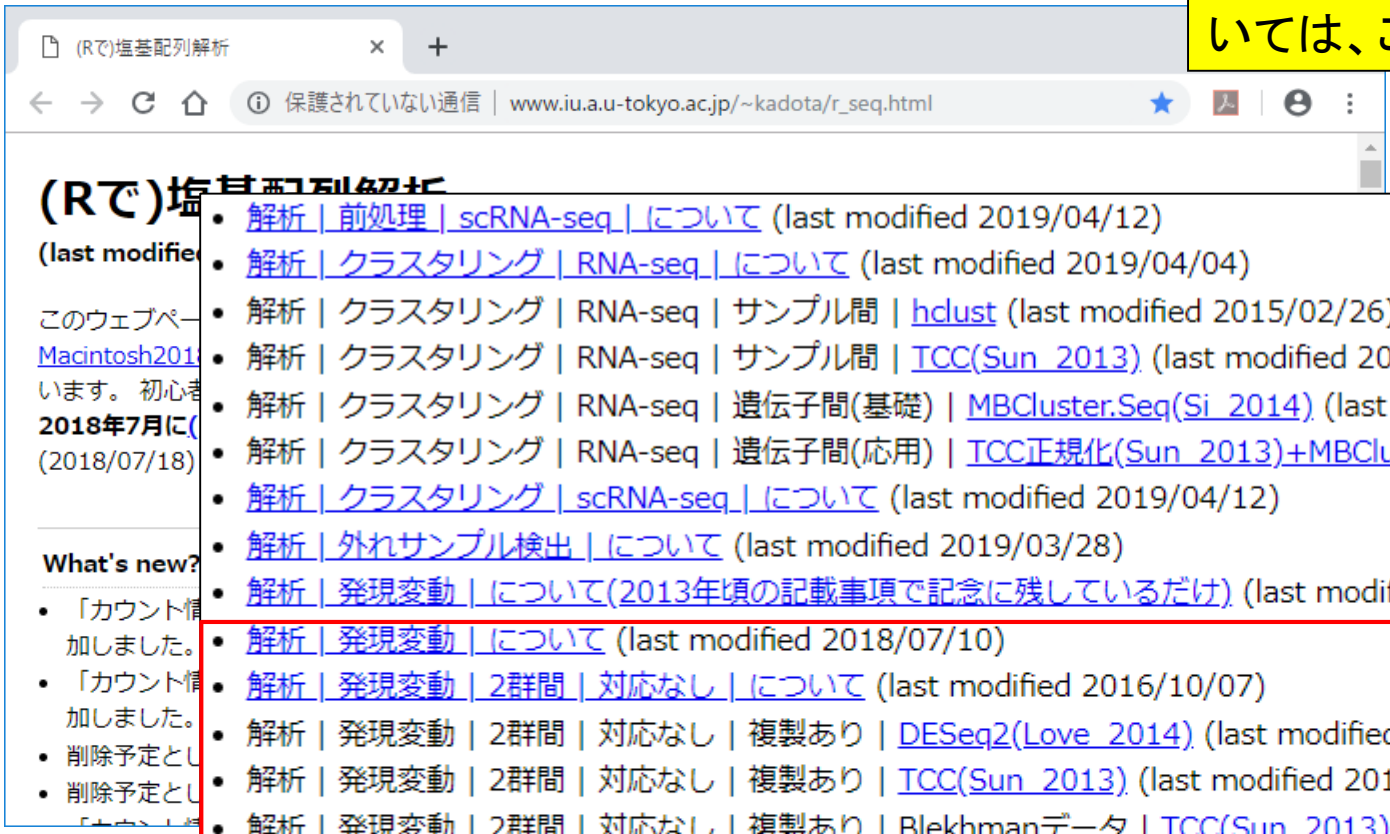
- 解析 | 基礎 | k-mer | ゲノムサイズ推定(基礎) | qrqc (last modified 2016/01/06)
- 解析 | 基礎 | 平均-分散プロット | について (last modified 2015/11/11)
- 解析 | 基礎 | 平均-分散プロット | Technical replicates (last modified 2014/02/18)
- 解析 | 基礎 | 平均-分散プロット | Biological replicates (last modified 2014/02/21)
- 解析 | 新規転写物同定(ゲノム配列を利用) (last modified 2019/05/21) NEW
- 解析 | 発現量推定(トランスクリプトーム配列を利用) (last modified 2019/04/05)
- 解析 | 解析 | 融合遺伝子の同定 (last modified 2019/05/21) NEW
- 解析 | 発現量推定(ゲノム配列を利用) (last modified 2016/10/04)

②

- 解析 | 前処理 | 型変換 | について (last modified 2019/04/03)
- 解析 | 前処理 | 型変換 | ExpressionSet --> SummarizedExperiment (last modified 2018/08/12)
- 解析 | 前処理 | 型変換 | ExpressionSet --> RangedSummarizedExperiment (last modified 2018/08/12)
- 解析 | 前処理 | 型変換 | RangedSummarizedExperiment --> ExpressionSet (last modified 2018/08/12)
- 解析 | 前処理 | 型変換 | SCESet --> CellDataSet (last modified 2019/04/03)
- 解析 | 前処理 | フィルタリング | 低発現遺伝子 | について (last modified 2019/03/28)
- 解析 | 前処理 | フィルタリング | 低発現遺伝子 | 基礎 (last modified 2018/08/08)
- 解析 | 前処理 | フィルタリング | 低発現遺伝子 | TCC(Sun 2013) (last modified 2018/08/08)

参考

①のページ上では、②のところに遺伝子構造推定、発現量推定、融合遺伝子検出の項目があります。③発現変動解析については、このあたりに沢山項目があります



- 解析 | 前処理 | scRNA-seq | について (last modified 2019/04/12)
- 解析 | クラスタリング | RNA-seq | について (last modified 2019/04/04)
- 解析 | クラスタリング | RNA-seq | サンプル間 | hclust (last modified 2015/02/26)
- 解析 | クラスタリング | RNA-seq | サンプル間 | TCC(Sun 2013) (last modified 2018/08/06)
- 解析 | クラスタリング | RNA-seq | 遺伝子間(基礎) | MBCluster.Seq(Si 2014) (last modified 2018/09/23)
- 解析 | クラスタリング | RNA-seq | 遺伝子間(応用) | TCC正規化(Sun 2013)+MBCluster.Seq(Si 2014) (last modified 2016/05/30)
- 解析 | クラスタリング | scRNA-seq | について (last modified 2019/04/12)
- 解析 | 外れサンプル検出 | について (last modified 2019/03/28)
- 解析 | 発現変動 | について(2013年頃の記載事項で記念に残しているだけ) (last modified 2014/07/10)
- 解析 | 発現変動 | について (last modified 2018/07/10)
- 解析 | 発現変動 | 2群間 | 対応なし | について (last modified 2016/10/07)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | DESeq2(Love 2014) (last modified 2015/11/15)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC(Sun 2013) (last modified 2015/07/07)推奨
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | Blekhmanデータ | TCC(Sun 2013) (last modified 2015/07/07)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | SAMseq(Li 2013) (last modified 2014/02/07)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | edgeR(Robinson 2010) (last modified 2014/07/24)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | WAD(Kadota 2008) (last modified 2015/03/30)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | varSelRF(Diaz-Uriarte 2007) (last modified 2019/02/10)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製なし | について (last modified 2018/07/23)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製なし | CORNAS(Low 2017) (last modified 2018/07/19)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製なし | LPEseq(Gim 2016) (last modified 2018/12/12)

③

Contents

- トランスクリプトーム解析技術の原理や特徴
  - RNA-seq（Illuminaの場合）、遺伝子≠転写物
- RNA-seqデータ解析のイメージ
  - マッピング→新規転写物の同定
- 様々な解析目的
  - 転写物配列取得、手法比較論文の紹介、ウェブツール（Maser）
  - データ解析の全体像（入出力の関係や代表的なツール）
- アノテーションファイルの読み込みと課題1
  - Rで転写物配列取得のイントロ
- Rで転写物配列取得と課題2
  - アノテーションファイルとゲノム情報ファイルから
- 公共データベース
  - NGS全体（NCBI SRA, EMBL-EBI ENA, DDBJ SRA）
  - DRAの概要、クオリティスコアなど

アノテーション

遺伝子構造（転写領域）推定結果ファイル≒アノテーションファイル。この形式は、GFF/GTFが有名。拡張子は.gffや.gtfや.gff3など。GFFはgeneral feature formatの略。GFF3は、GFF ver3という意味。GTFはGFF ver2.5とも呼ばれる。

- イントロ | NGS | 配列取得 | FASTQ or SRA | SRAdb(Zhu 2013) (last modified 201
- イントロ | NGS | 配列取得 | シミュレーションデータ | について (last modified 2015/
- イントロ | NGS | 配列取得 | シミュレーションデータ | ランダムな塩基配列の生成か
- イントロ | NGS | アノテーション情報取得 | について (last modified 2014/03/26)
- イントロ | NGS | アノテーション情報取得 | GFF/GTF形式ファイル (last modified 2017/04/10) NEW
- イントロ | NGS | アノテーション情報取得 | refFlat形式ファイル (last modified 2013/09/25)
- イントロ | NGS | アノテーション情報取得 | biomaRt(Durinck 2009) (last modified 2013/09/26)
- イントロ | NGS | アノテーション情報取得 |
- イントロ | NGS | アノテーション情報取得 |
- イントロ | NGS | アノテーション情報取得 |
- イントロ | NGS | アノテーション情報取得 |
- イントロ | NGS | 読み込み | BSgenome |
- イントロ | NGS | 読み込み | FASTA形式
- イントロ | NGS | 読み込み | FASTA形式
- イントロ | NGS | 読み込み | FASTQ形式
- イントロ | NGS | 読み込み | FASTQ形式
- イントロ | NGS | 読み込み | FASTQ形式
- イントロ | NGS | 読み込み | SAM/BAM形

① ②

イントロ | NGS | アノテーション情報取得 | GFF/GTF形式ファイル NEW

多くの生物種についてEnsembl (Yates et al., Nucleic Acids Res., 2016) のFTPサイトから GTF形式(GFF ver. 2)の遺伝子アノテーションファイルを得ることができます。2017年4月10日現在、上記のFTPサイトへのアクセス時にユーザ名とパスワードを聞かれますが、「匿名でログオンする」にチェックを入れればOKです。GTFはGeneral Transfer FormatまたはGene Transfer Formatの略で、GTFの派生版としてGTF2というフォーマットもあるようです。また、General Feature Format ver. 3 (GFF3)という形式も存在するなど、GFF/GTF形式として総称されている中で様々なバリエーションがあります。いずれもrefFlat形式同様、どの領域にどの遺伝子があるのかという座標(Coordinates)情報を含みます。ゲノム配列のバージョンと同じであることを確認した上で用いましょう。

- Ensembl (Yates et al., Nucleic Acids Res., 2016)
  圧縮(gzip)ファイル形式です。基本はFTPサイトです。代表的なものを以下にリストアップしています。
  - ヒト; Human (H.sapiens)
  - ラット; Rat (R.norvegicus)
  - ネコ; Cat (F.catus)
  - ウサギ; Rabbit (O.cuniculus)
  - ニワトリ; Chicken (G.gallus)
  - イヌ; Dog (C.familiaris)
  - ウマ; Horse (E.caballus)
  - ゼブラフィッシュ; Zebrafish (D.rerio)
  - ...
- イネ: RAP-DB (Sakai et al., Plant Cell Physiol., 2013)
  - 「ダウンロード」-「Gene set」-「Gene structure and function information in GFF format」-「Download」。IRGSP-1.0_representative_2014-03-05.tar.gz (12.4MB程度)の圧縮ファイルが得られます。
- シロイヌナズナ: The Arabidopsis Information Resource (TAIR) (Lamesch et al., Nucleic Acids Res., 2012)
  - 「ダウンロード」-「Genes」-「TAIR10_genome_release」-「TAIR10_gff3」の TAIR10_GFF3_genes.gff (42MB程度)

他にrefFlat形式など様々なファイル形式が存在します

GFF/GTF形式ファイルの例

GFF3形式（シロイヌナズナ; TAIR10_GFF3_genes.gff）

	A	B	C	D	E	F	G	H	I
1	Chr1	TAIR10	chromosome	1	30427671	.	.	.	ID=Chr1;Name=Chr1
2	Chr1	TAIR10	gene	3631	5899	.	+	.	ID=AT1G01010;Note=protein_coding_gene;Name=AT1G01010
3	Chr1	TAIR10	mRNA	3631	5899	.	+	.	ID=AT1G01010.1;Parent=AT1G01010;Name=AT1G01010.1;Index=1
4	Chr1	TAIR10	protein	3760	5630	.	+	.	ID=AT1G01010.1-Protein;Name=AT1G01010.1;Derives_from=AT1G01010.1
5	Chr1	TAIR10	exon	3631	3913	.	+	.	Parent=AT1G01010.1
6	Chr1	TAIR10	five_prime_UTR	3631	3759	.	+	.	Parent=AT1G01010.1
7	Chr1	TAIR10	CDS	3760	3913	.	+	0	Parent=AT1G01010.1,AT1G01010.1-Protein;
8	Chr1	TAIR10	exon	3996	4276	.	+	.	Parent=AT1G01010.1
9	Chr1	TAIR10	CDS	3996	4276	.	+	2	Parent=AT1G01010.1,AT1G01010.1-Protein;
10	Chr1	TAIR10	exon	4486	4605	.	+	.	Parent=AT1G01010.1
11	Chr1	TAIR10	CDS	4486	4605	.	+	0	Parent=AT1G01010.1,AT1G01010.1-Protein;
12	Chr1	TAIR10	exon	4706	5095	.	+	.	Parent=AT1G01010.1

GTF形式（ゼブラフィッシュ; Danio_rerio.Zv9.75.gtf）

	A	B	C	D	E	F	G	H	I
1	#!genome-build Zv9								
2	#!genome-version Zv9								
3	#!genome-date 2010-04								
4	#!genome-build-accession NCBI:GCA_000002035.2								
5	#!genebuild-last-updated 2014-02								
6	7	protein_coding	gene	100958	101715	.	+	.	gene_id "ENSDARG00000076051"; gene_name "CABZ01062994.1"; gene
7	7	protein_coding	transcript	100958	101715	.	+	.	gene_id "ENSDARG00000076051"; transcript_id "ENSDART00000113409
8	7	protein_coding	exon	100958	100975	.	+	.	gene_id "ENSDARG00000076051"; transcript_id "ENSDART00000113409
9	7	protein_coding	CDS	100958	100975	.	+	0	gene_id "ENSDARG00000076051"; transcript_id "ENSDART00000113409
10	7	protein_coding	exon	101077	101715	.	+	.	gene_id "ENSDARG00000076051"; transcript_id "ENSDART00000113409
11	7	protein_coding	CDS	101077	101715	.	+	0	gene_id "ENSDARG00000076051"; transcript_id "ENSDART00000113409
12	7	protein_coding	gene	116160	117573	.	+	.	gene_id "ENSDARG00000088691"; gene_name "BX511027.1"; gene_sour
13	7	protein_coding	transcript	116160	117573	.	+	.	gene_id "ENSDARG00000088691"; transcript_id "ENSDART00000129330

GFFの読み込み

読み込み段階でコケる、読み込みはうまくいったが、その後の解析段階でコケるなど、Linux上での解析同様、一筋縄ではいきません。過去の受講生など多方面からの情報提供のおかげでだいぶ分かってきました。

①

イントロ | NGS | アノテーション情報取得 | TxDb | GFF/GTF形式ファイルから NEW

QuasRパッケージを用いてゲノムへのマッピング結果からカウント情報を得たいときに、"TxDb"という形式のオブジェクトを利用する必要があります。ここでは、GenomicFeatures パッケージを用いて手元にあるGFF/GTF形式ファイルを入力としてTxDbオブジェクトを得るやり方を示します。基本的にはGenomicFeaturesパッケージ中のmakeTxDbFromGFF関数を用いてGFF/GTF形式ファイルを読み込むことで TxDbオブジェクトをエラーなく読み込むこと自体は簡単にできます。しかし、得られたTxDbオブジェクトとゲノムマッピング結果ファイルを用いてカウント情報を得る場合に、ゲノム配列提供元とアノテーション情報提供元が異なっているとエラーとなります。具体的には、GFF/GTFファイル中にゲノム配列中にない染色体名があるとエラーが出る場合があります。

1. TAIR(Lamesch et al., Nucleic Acids Res., 2012) から提供されているArabidopsisのGFF3形式ファイル(TAIR10_GFF3_genes.gff)の場合:

基本形です。エラーは出ませんが、2015年3月4日現在、ChrCが環状ではないと認識されてしまっています。

```
in_f <- "TAIR10_GFF3_genes.gff"          #入力ファイル名を指定してin_fに格納(GFF/GTFファイル)

#必要なパッケージをロード
library(GenomicFeatures)                 #パッケージの読み込み

#本番(TxDbオブジェクトの作成)
txdb <- makeTxDbFromGFF(in_f)            #txdbオブジェクトの作成
txdb                                     #確認してるだけです
```

GFFの読み込み

①例題7。②ここで用いているGFF形式の入力ファイルは、③から取得しました。③をクリックしたつもりで次のスライドを眺める。デスクトップ上のhogeフォルダ内に②のgffファイルはあります

イントロ | NGS | アノテーション情報取得 | TxDb | GFF/GTF形式ファイルから NEW

QuasRパッケージを用いてゲノムへのマッピング結果からカウント情報を得たいときに、"TxDb"という形式のオブジェクトを利用する必要があります。ここでは、GenomicFeatures パッケージを用いて手元にあるGFF/GTF形式ファイルを入力としてTxDbオブジェクトを得るやり方を示します。基本的にはGenomicFeaturesパッケージ中のmakeTxDbFromGFF関数を用いてGFF/GTF形式ファイルを読み込むことで TxDbオブジェクトをエラーなく読み込むこと自体は簡単にできます。しかし、得られたTxDbオブジェクトとゲノムマッピング結果ファイルを用いてカウント情報を得る場合に、ゲノム配列提供元とアノテーション情報提供元が異なっているとエラーとなります。具体的には、GFF/GTFファイル中にゲノム配列中にない染色体名があるとエラーが出る場合があります。

1. TAIR(Lamesch et al., Nucleic Acids Res., 2012) から提供されているArabidopsisのGFF3形式ファイル(TAIR10_GFF3_genes.gff)の場合:

基本形です。エラーは出ませんが、2015年3月4日現在、ChrCが環状ではないと認識されてしまっています。

```
in_f <- "TAIR10_GFF3_genes.gff"          #入力ファイル名を指定してin_fに格納(GFF/GTFファイル)

#必要
libr
#本番(T
txdb <-
txdb
```

①

7. GFF3形式ファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.gff3)の場合:

Ensembl (Flicek et al., 2014)から提供されている Lactobacillus hokkaidonensis JCM 18461 (Tanizawa et al., 2015) のデータです。

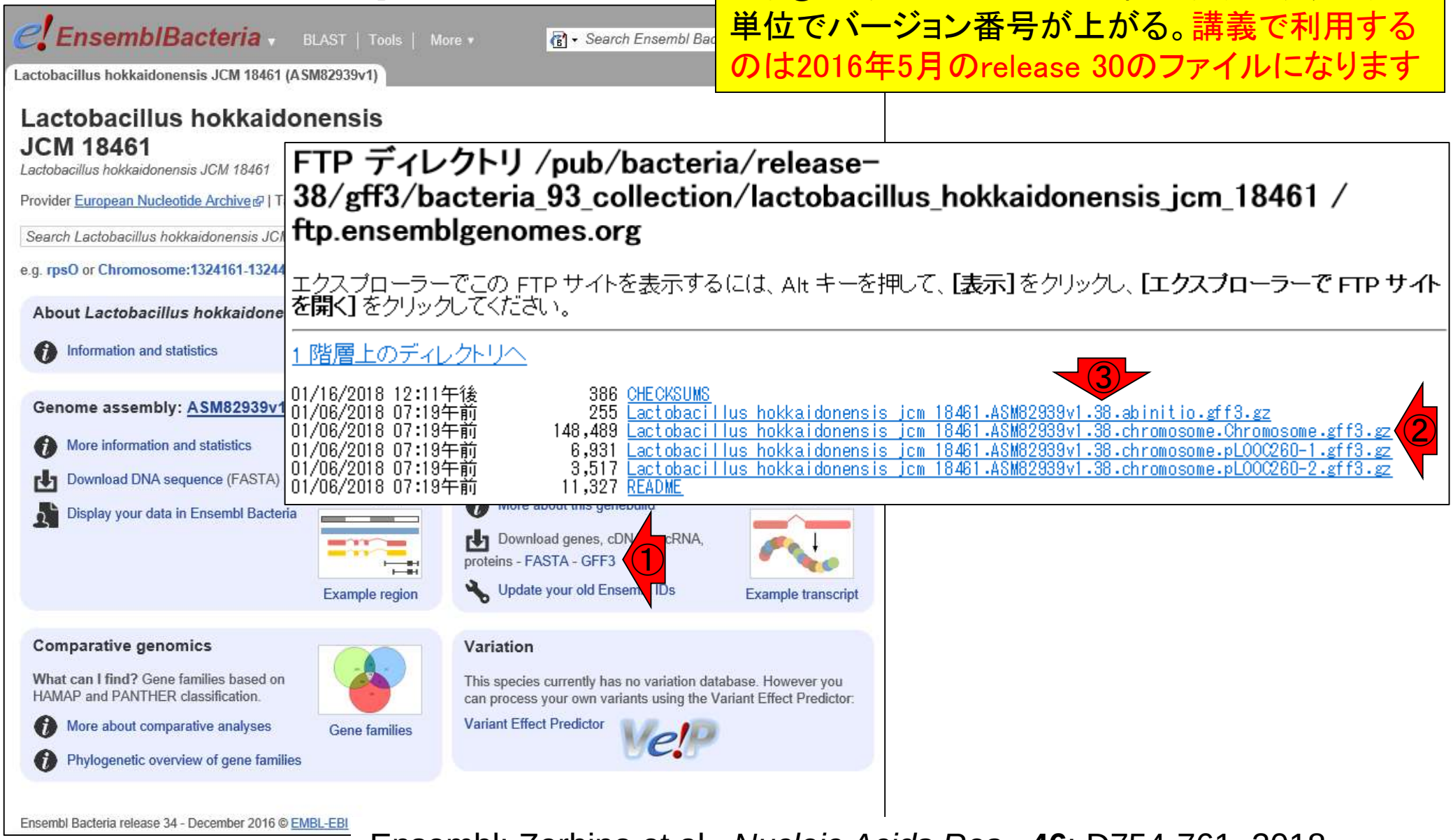
③

```
in_f <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.gff3"#入

#必要なパッケージをロード
library(GenomicFeatures)                  #パッケージの読み込み

#本番(TxDbオブジェクトの作成)
txdb <- makeTxDbFromGFF(in_f, format="auto")#txdbオブジェクトの作成
txdb                                      #確認してるだけです
```

②

Ensembl解説

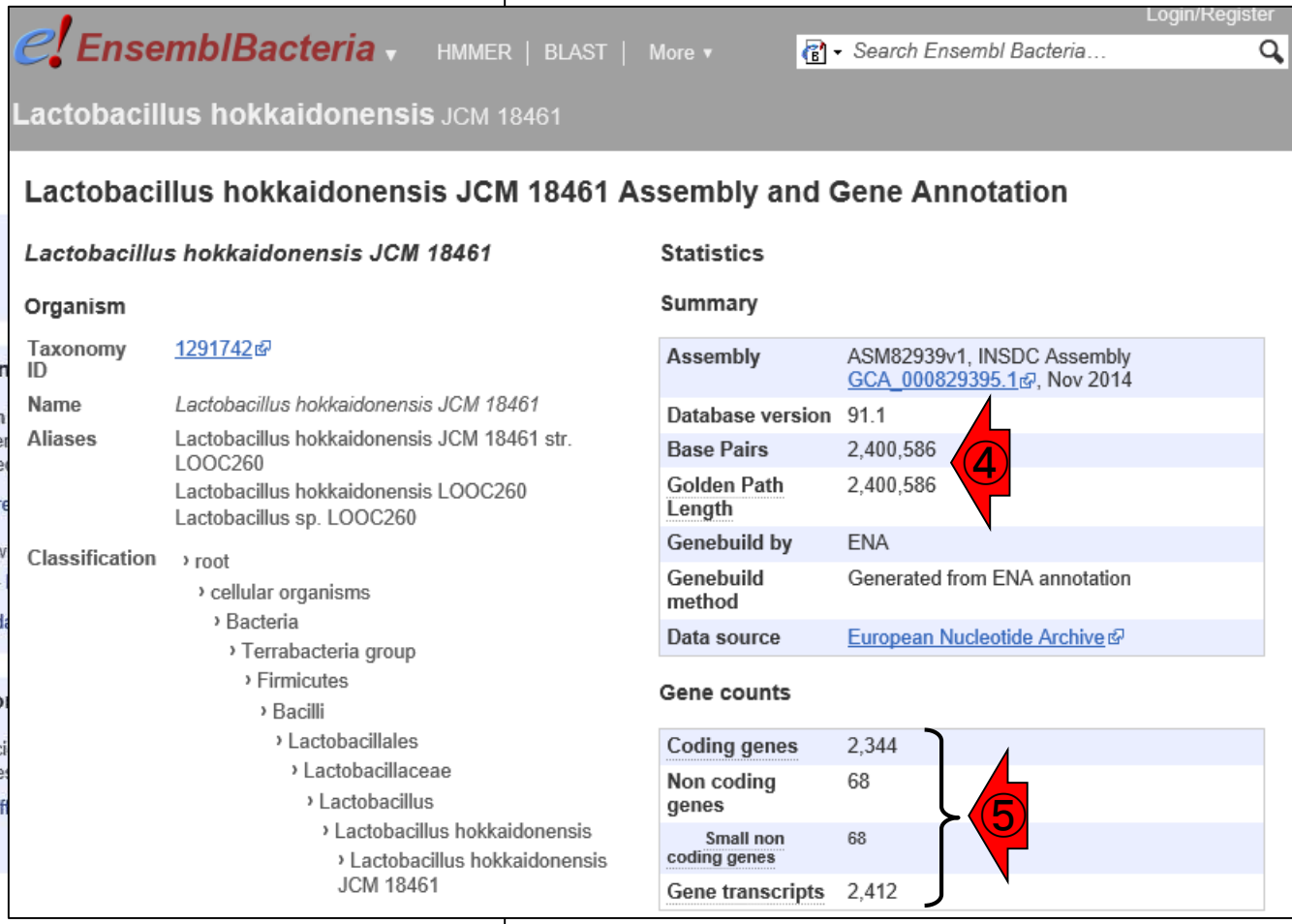
①GFFファイルはここから取得。②のgzip圧縮ファイルをダウンロードして解凍したものが入力ファイル。③のあたりがバージョン番号。概ね、数か月単位でバージョン番号が上がる。講義で利用するのは2016年5月のrelease 30のファイルになります



Ensembl: Zerbino et al., *Nucleic Acids Res.*, **46**: D754-761, 2018

Ensembl解説

①このゲノムの全貌は②である程度把握可能。③原著論文の情報なども合わせることで、④ゲノムサイズが約2.4MB、⑤2,344 coding genesなどの情報がわかる。⑥でゲノム配列をダウンロードできる



③Tanizawa et al., *BMC Genomics*, **16**: 240, 2015

Ensembl解説

いろんなものがあって私はよくわかりませんが、GFFファイルと一緒に取り扱いたいときには、GFFファイルと似た名前の①を採用します。正確には、このゲノムは1つの染色体と②2つのプラスミド(pLOOC260-1とpLOOC260-2)からなっています。①はそのうちの染色体配列のみになります。③ファイルサイズ的に、これが3つの配列がまとめられたものなのでしょう

FTP ディレクトリ /pub/bacteria/releas
38/fasta/bacteria_93_collection/lactol
ftp.ensemblgenomes.org

エクスプローラーでこの FTP サイトを表示するには、Alt キー を押して、[表示] をクリックし、[エクスプローラー で FTP サイトを開く] をクリックしてください。

1 階層上のディレクトリへ

01/15/2018 09:56午後	1,132	CHECKSUMS
01/08/2018 12:24午後	706,219	Lactobacillus hokkaidonensis jcm 18461.ASM82939v1.dna.chromosome.Chromosome.fa.gz
01/08/2018 12:24午後	26,014	Lactobacillus hokkaidonensis jcm 18461.ASM82939v1.dna.chromosome.pLOOC260-1.fa.gz
01/08/2018 12:24午後	13,347	Lactobacillus hokkaidonensis jcm 18461.ASM82939v1.dna.chromosome.pLOOC260-2.fa.gz
01/08/2018 12:24午後	745,580	Lactobacillus hokkaidonensis jcm 18461.ASM82939v1.dna.toplevel.fa.gz
01/08/2018 12:24午後	706,228	Lactobacillus hokkaidonensis jcm 18461.ASM82939v1.dna_rm.chromosome.[illegible]osome.fa.gz
01/08/2018 12:24午後	26,023	Lactobacillus hokkaidonensis jcm 18461.ASM82939v1.dna_rm.chromosome.[illegible]OOC260-1.fa.gz
01/08/2018 12:24午後	13,355	Lactobacillus hokkaidonensis jcm 18461.ASM82939v1.dna_rm.chromosome.pLOOC260-2.fa.gz
01/08/2018 12:24午後	745,606	Lactobacillus hokkaidonensis jcm 18461.ASM82939v1.dna_rm.toplevel.fa.gz
01/08/2018 12:24午後	706,228	Lactobacillus hokkaidonensis jcm 18461.ASM82939v1.dna_sm.chromosome.Chromosome.fa.gz
01/08/2018 12:24午後	26,022	Lactobacillus hokkaidonensis jcm 18461.ASM82939v1.dna_sm.chromosome.pLOOC260-1.fa.gz
01/08/2018 12:24午後	13,355	Lactobacillus hokkaidonensis jcm 18461.ASM82939v1.dna_sm.chromosome.pLOOC260-2.fa.gz
01/08/2018 12:24午後	745,605	Lactobacillus hokkaidonensis jcm 18461.ASM82939v1.dna_sm.toplevel.fa.gz
01/08/2018 12:24午後	4,923	README

GFFの読み込み

①例題7が読み込みの基本形。②GenomicFeaturesというパッケージが提供する③makeTxDbFromGFF関数を用いてGFFファイルを読み込んで、TxDbという独特の形式で取り扱えるようにする

7. GFF3形式ファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829

Ensembl (Flicek et al., 2014)から提供されている Lactobacillus hokkaidonensis JCM 18461 (Tanizawa et al., 2015)のデータです。

```
in_f <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.gff3"#入

#必要なパッケージをロード
library(GenomicFeatures)                #パッケージの読み込み

#本番(TxDbオブジェクトの作成)
txdb <- makeTxDbFromGFF(in_f, format="auto")#txdbオブジェクトの作成
txdb                                    #確認してるだけです
```

GFFの読み込み

①読み込み後のtxdbオブジェクトの中身を表示

7. GFF3形式ファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.gff3)の場合:

Ensembl (Flicek et al., 2014)から提供されている Lactobacillus hokkaidonensis JCM 18461 (Tanizawa et al., 2015) のデータです。

```
in_f <- "Lactobacillus_ho

#必要なパッケージをロード
library(GenomicFeatures)

#本番(TxDbオブジェクトの作
txdb <- makeTxDbFromGFF(
txdb
```

①

R Console

```
> txdb <- makeTxDbFromGFF(in_f, format="auto")#txdbオブジェクトの作成
Import genomic features from the file as a GRanges object ... OK
Prepare the 'metadata' data frame ... OK
Make the TxDb object ... OK
> txdb                                          #確認してるだけです
TxDb object:
# Db type: TxDb
# Supporting package: GenomicFeatures
# Data source: Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.$
# Organism: NA
# Taxonomy ID: NA
# miRBase build ID: NA
# Genome: NA
# transcript_nrow: 2262
# exon_nrow: 2262
# cds_nrow: 2194
# Db created by: GenomicFeatures package from Bioconductor
# Creation time: 2018-04-13 16:12:15 +0900 (Fri, 13 Apr 2018)
# GenomicFeatures version at creation time: 1.30.3
# RSQLite version at creation time: 2.1.0
# DBSCHEMAVERSION: 1.2
>
```

矛盾?!

このゲノムは、1つの染色体と2つのプラスミド(pLOOC260-1とpLOOC260-2)からなっています。
①の結果は染色体のみの数値です。②のEnsemblウェブサイト上で見られる数値と一致していません

7. GFF3形式ファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_0008293

Ensembl (Flicek et al., 2014)から提供されている Lactobacillus hokkaidonensis JCM 18461 (Tanizawa et al

```
in_f <- "Lactobacillus_ho

#必要なパッケージをロード
library(GenomicFeatures)

#本番(TxDbオブジェクトの作
txdb <- makeTxDbFromGFF(
txdb
```

R Console

```
> txdb <- makeTxDbFromGFF(in_f, format=
Import genomic features from the file a
Prepare the 'metadata' data frame ... O
Make the TxDb object ... OK
> txdb
TxDb object:
# Db type: TxDb
# Supporting package: GenomicFeatures
# Data source: Lactobacillus_hokkaidone
# Organism: NA
# Taxonomy ID: NA
# miRBase build ID: NA
# Genome: NA
# transcript_nrow: 2262
# exon_nrow: 2262
# cds_nrow: 2194
# Db created by: GenomicFeatures packag
# Creation time: 2018-04-13 16:12:15 +0
# GenomicFeatures version at creation t
# RSQLite version at creation time: 2.1.0
# DBSCHEMAVERSION: 1.2
>
```

①

Assembly	ASM82939v1, INSDC Assembly GCA_000829395.1, Nov 2014
Database version	87.1
Base Pairs	2,400,586
Golden Path Length	2,400,586
Genebuild by	ENA
Genebuild method	Generated from ENA annotation
Data source	European Nucleotide Archive

Gene counts

Coding genes	2,344
Non coding genes	68
Small non coding genes	68
Gene transcripts	2,412

②

課題1

このゲノムは、1つの染色体と2つのプラスミド(pLOOC260-1とpLOOC260-2)からなっています。①の結果は染色体のみの数値です。②のEnsemblウェブサイト上で見られる数値と一致していません。プラスミドのgffファイル(plasmid1.gff3とplasmid2.gff3)をそれぞれ読み込んで①transcript_nrow (Gene transcripts)と②cds_nrow (Coding genes)の情報を得て、③Ensemblウェブサイト上の数値と絡めて簡単に考察せよ

7. GFF3形式ファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_0008293

Ensembl (Flicek et al., 2014)から提供されている Lactobacillus hokkaidonensis JC

```
in_f <- "Lactobacillus_h

#必要なパッケージをロード
library(GenomicFeatures)

#本番(TxDbオブジェクトの作
txdb <- makeTxDbFromGFF(
txdb
```

R Console

```
> txdb <- makeTxDbFromGFE
Import genomic features f
Prepare the 'metadata' da
Make the TxDb object ...
> txdb
TxDb object:
# Db type: TxDb
# Supporting package: GenomicFeatures
# Data source: Lactobacillus_hokkaidone
# Organism: NA
# Taxonomy ID: NA
# miRBase build ID: NA
# Genome: NA
① transcript_nrow: 2262
# exon_nrow: 2262
② cds_nrow: 2194
# Db created by: GenomicFeatures packag
# Creation time: 2018-04-13 16:12:15 +0
# GenomicFeatures version at creation t
# RSQLite version at creation time: 2.1.0
# DBSCHEMAVERSION: 1.2
>
```

Length	
Genebuild by	ENA
Genebuild method	Generated from ENA annotation
Data source	European Nucleotide Archive

Gene counts

Coding genes	2,344 ③
Non coding genes	68
Small non coding genes	68
Gene transcripts	2,412 ③

課題1

①プラスミドのgffファイル（plasmid1.gff3とplasmid2.gff3）はこちら

9.機能ゲノム学

授業の目標・概要

細胞中で発現している全転写物（トランスクリプトーム）解析手法について、特に塩基配列解析部分を中心に解説します。また、Rのスキルアップを目指します。

担当教員

門田幸二 （東大・農・アグリバイオ / 准教授）

お知らせ

講義では、Rの様々なパッケージを利用します。持ち込み用PC利用希望者は インストール | について を参考にしてR本体および必要なパッケージ群を必ずインストールしておいてください。

フリーソフトウェアRの基本的な利用法を習得済みであることを前提として行いますので、 基本的な利用方法 を参考にして基礎的な事柄を理解しておいてください。

参考図書

門田幸二 著 (金明哲 編)、「シリーズ Useful R ⑦ トランスクリプトーム解析」、共立出版、2014。ISBN:978-4-320-12370-0

坊農秀雄 著、生命科学データ解析、MEDSi、2017

講義日程（2019年度）

1. 2019年05月27日
   - 講義資料PDF
   - .gff3ファイル (約1.3MB)
   - .faファイル (約2.2MB)
   - (Rで)塩基配列解析
   - (Rで)マイクロアレイデータ解析
   - plasmid1.gff3(課題用)
   - plasmid2.gff3(課題用)
   - de Lannoy et al., F1000Res., 2017
   - Garalde et al., Nat Methods, 2018
   - RNACocktail : Sahraeian et al., Nat Commun., 2017
2. 2019年06月03日
3. 2019年06月10日
4. 2019年06月17日

Contents

- トランスクリプトーム解析技術の原理や特徴
  - RNA-seq（Illuminaの場合）、遺伝子≠転写物
- RNA-seqデータ解析のイメージ
  - マッピング→新規転写物の同定
- 様々な解析目的
  - 転写物配列取得、手法比較論文の紹介、ウェブツール（Maser）
  - データ解析の全体像（入出力の関係や代表的なツール）
- アノテーションファイルの読み込みと課題1
  - Rで転写物配列取得のイントロ
- Rで転写物配列取得と課題2
  - アノテーションファイルとゲノム情報ファイルから
- 公共データベース
  - NGS全体（NCBI SRA, EMBL-EBI ENA, DDBJ SRA）
  - DRAの概要、クオリティスコアなど

転写物配列取得

multi-FASTAファイル（ゲノム配列情報）とGFFファイル（アノテーション情報）を同時に読み込むことで、①**トランスクリプトーム（転写物）配列情報を一気に取得することも可能**。②例題5。③はhogeフォルダ中にあります

- イントロ | 一般 | 配列取得 | プロモーター配列 | BSgenomeとTxDbから(last
- イントロ | 一般 | 配列取得 | プロモーター配列 | GenomicFeatures(Lawrenc
- イントロ | 一般 | 配列取得 | トランスクリプトーム配列 | 公共DBから(last modified 2015/05/
- イントロ | 一般 | 配列取得 | トランスクリプトーム配列 | GenomicFeatures(Lawrence_2013)(last modified 2016/02/09)
- イントロ | 一般 | 配列取得 | トランスクリプトーム配列 | biomaRt(Durinck_2009)(last modifie 2015/02/20)
- イントロ | 一般 | 読み込み | xlsx形式 | openxlsx(last modified 2015/11/15)
- イントロ | NG
- イントロ | NG
- イントロ | NG
- イントロ | NG
- イントロ | NG
- イントロ | NG
- イントロ | NG
- イントロ | NG

①

イントロ | 一般 | 配列取得 | トランスクリプトーム配列 | GenomicFeatures(Lawrence_2013)

NEW

GenomicFeaturesパッケージを主に用いてトランスクリプトーム配列を得るやり方を示します。「?extractTranscriptSeqs」を行うことで、様々な例題を見ることができます。transcriptsBy関数部分は、exonsBy, cdsBy, intronsByTranscript, fiveUTRsByTranscript, threeUTRsByTranscriptなど様々な他の関数で置き換えることができます。

②

「ファイル

1. ヒト (BSg

対応するア
果は82,960

```
out_f <-
param_bsg
param_txd

#必要なパ
library(G

#前処理(ゲ
library(p
tmp <- ls
genome <-
```

5. GFF3形式のアノテーションファイルとFASTA形式のゲノム配列ファイルを読み込む場合:

GFF3形式ファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.gff3)とFASTA形式ファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa)を読み込むやり方です。Ensembl (Flicek et al., 2014)から提供されている Lactobacillus hokkaidonensis JCM 18461 (Tanizawa et al., 2015) のデータです。

```
in_f1 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa"
in_f2 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.gff3"#入
out_f <- "hoge5.fasta"                    #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(Rsamtools)                        #パッケ
library(GenomicFeatures)                  #パッケ
library(Biostrings)                       #パッケ

#入力ファイルの読み込み
txdb <- makeTxDbFromGFF(in_f2, format="auto")#
txdb                                      #確認し

#前処理(欲しい領域の座標情報取得)
hoge <- transcripts(txdb)                 #指定し
hoge                                      #確認し

#本番(配列取得)
```

③

1. 2019年05月27日
   - 講義資料PDF
   - .gff3ファイル (約1.3MB)
   - .faファイル (約2.2MB)
   - (Rで)塩基配列解析
   - (Rで)マイクロアレイデータ解析
   - plasmid1.gff3(課題用)
   - plasmid2.gff3(課題用)
   - de Lannoy et al., F1000Res., 2017
   - Garalde et al., Nat Methods, 2018
   - RNACocktail : Sahraeian et al., Nat Commun., 2017

転写物配列取得

①は、GFFファイル情報を保持したtxdbオブジェクトから、transcriptsという関数を用いて抽出したい転写物の座標情報を取得した結果を②hogeに保存している

5. GFF3形式のアノテーションファイルとFASTA形式のゲノム配列ファイルを読み込む場合:

GFF3形式ファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.gff3)とFASTA形式ファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa)を読み込むやり方です。
Ensembl (Flicek et al., 2014)から提供されている Lactobacillus hokkaidonensis JCM 18461 (Tanizawa et al., 2015) のデータです。

```
in_f1 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa"#ノ
in_f2 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.gff3"#入力
out_f <- "hoge5.fasta"                    #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(Rsamtools)
library(GenomicFeatures)
library(Biostrings)

#入力ファイルの読み込み
txdb <- makeTxDbFromGFF(in_
txdb

#前処理(欲しい領域の座標情報取
hoge <- transcripts(txdb)
hoge

#本番(配列取得)
fasta <- getSeq(FaFile(in_f
fasta

#後処理(description部分を変更
```

② ②

R Console

```
> #前処理(欲しい領域の座標情報取得)
> hoge <- transcripts(txdb)                    #指定した範囲の座標情報を取得
> hoge                                         #確認してるだけです
GRanges object with 2262 ranges and 2 metadata columns:
          seqnames              ranges strand   |     tx_id     tx_name
             <Rle>           <IRanges>  <Rle>   | <integer> <character>
      [1] Chromosome       [ 360, 1676]      +   |         1       dnaA-1
      [2] Chromosome       [1852, 2991]      +   |         2       dnaN-1
      [3] Chromosome       [3233, 3457]      +   |         3         <NA>
      [4] Chromosome       [3467, 4588]      +   |         4       recF-1
      [5] Chromosome       [4588, 6531]      +   |         5       gyrB-1
      ...           ...                ...    ... ...       ...          ...
   [2258] Chromosome [2273924, 2275312]      -   |      2258       trmE-1
   [2259] Chromosome [2275488, 2276288]      -   |      2259         <NA>
   [2260] Chromosome [2276455, 2277288]      -   |      2260         <NA>
   [2261] Chromosome [2277304, 2277648]      -   |      2261         <NA>
   [2262] Chromosome [2277719, 2277853]      -   |      2262       rpmH-1
   -------
   seqinfo: 1 sequence from an unspecified genome; no seqlengths
>
```

① ①

転写物配列取得

GFFファイルの見方がよくわかっていなくても、GFFファイル中の①のあたりとhogeオブジェクト中の②と比較することで、うまく読み込めているらしいことはわかる

```
##gff-version  3
##sequence-region   Chromosome 360 2277853
#!genome-build European Nucleotide Archive ASM82939
#!genome-version GCA_000829395.1
#!genome-date 2014-11
#!genome-build-accession GCA_000829395.1
#!genebuild-last-updated 2014-11
Chromosome  ena  gene        360  1676  .  +  .  ID=ge
Chromosome  ena  transcript  360  1676  .  +  .  ID=tr
Chromosome  ena  exon        360  1676  .  +  .  Pare
Chromosome  ena  CDS         360  1676  .  +  0  ID=C
###
Chromosome  ena  gene        1852 2991  .  +  .  ID=ge
Chromosome  ena  transcript  1852 2991  .  +  .  ID=tr
Chromosome  ena  exon        1852 2991  .  +  .  Pare
Chromosome  ena  CDS         1852 2991  .  +  0  ID=C
###
Chromosome  ena  gene        3233 3457  .  +  .  ID=ge
Chromosome  ena  transcript  3233 3457  .  +  .  ID=tr
Chromosome  ena  exon        3233 3457  .  +  .  Pare
Chromosome  ena  CDS         3233 3457  .  +  0  ID=C
###
Chromosome  ena  gene        3467 4588  .  +  .  ID=ge
```

①

```
標情報取得)
ts(txdb)                    #指定した範囲の座標情報を取得
                            #確認してるだけです
2262 ranges and 2 metadata columns:
            ranges strand   |      tx_id     tx_name
         <IRanges>  <Rle>   | <integer> <character>
      [ 360, 1676]      +   |         1       dnaA-1
      [1852, 2991]      +   |         2       dnaN-1
      [3233, 3457]      +   |         3         <NA>
      [3467, 4588]      +   |         4       recF-1
      [4588, 6531]      +   |         5       gyrB-1
               ...    ... ...       ...          ...
[2273924, 2275312]      -   |      2258       trmE-1
[2275488, 2276288]      -   |      2259         <NA>
[2276455, 2277288]      -   |      2260         <NA>
[2277304, 2277648]      -   |      2261         <NA>
[2277719, 2277853]      -   |      2262       rpmH-1
  -------
  seqinfo: 1 sequence from an unspecified genome; no seqlengths
> |
```

②

転写物配列取得

①in_f1で指定したゲノム配列情報は②ここで登場。①のゲノム配列から、③のhogeで指定した座標の塩基配列を④(Biostringsパッケージが提供する)getSeq関数を用いて取得。⑤(Rsamtoolsパッケージが提供する)FaFile関数は、getSeq関数利用時に必要

5. GFF3形式のアノテーションファイルとFASTA形式のゲノム配列ファイルを

GFF3形式ファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_0008293
ファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna
Ensembl (Flicek et al., 2014)から提供されている Lactobacillus hokkaidonensis JCM 18461 (Tanizawa et al., 2015)のデータです。

```
in_f1 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa"
in_f2 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.gff3"#入
out_f <- "hoge5.fasta"                    #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(Rsamtools)                        #パッケージの読み込み
library(GenomicFeatures)                  #パッケージの読み込み
library(Biostrings)                       #パッケージの読み込み

#入力ファイルの読み込み
txdb <- makeTxDbFromGFF(in_f2, format="auto")#txdbオブジェクトの作成
txdb                                      #確認してるだけです

#前処理(欲しい領域の座標情報取得)
hoge <- transcripts(txdb)                 #指定した範囲の座標情
hoge                                      #確認してるだけです

#本番(配列取得)
fasta <- getSeq(FaFile(in_f1), hoge)      #配列情報を取得した結果
fasta                                     #確認してるだけです

#後処理(description部分を変更)
```

①

④

⑤

④ ② ③

R Console

```
> FaFile(in_f1)
class: FaFile
path: .../Lactobacillus_hokkaidonensis_jcm_$
index: .../Lactobacillus_hokkaidonensis_jcm$
isOpen: FALSE
yieldSize: NA
>
```

転写物配列取得

①getSeq実行後の②fastaオブジェクトが、欲しいトランスクリプトーム配列情報ではあるが…

5. GFF3形式のアノテーションファイルとFASTA形式のゲノム配列ファイルを読み込む場合:

GFF3形式ファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.gff3)とFASTA形式ファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa)を読み込むやり方です。
Ensembl (Flicek et al., 2014)から提供されている Lactobacillus hokkaidonensis JCM 18461 (Tanizawa et al., 2015) のデータです。

```
library(Biostrings)                          #パッケージの読み込み

#入力ファイルの読み込み
txdb <- makeTxDbFromGFF(in_f2, format="auto")#txdbオブジェクトの作成
txdb                                         #確認してるだけです

#前処理(欲しい領域の座標情報取得)
hoge <- transcripts(txdb)
hoge

#本番(配列取得)
fasta <- getSeq(FaFile(in_f1), hoge)
fasta

#後処理(description部分を変更)
names(fasta) <- paste(seqnames(hoge),
                      end(ranges(hoge
fasta

#ファイルに保存
writeXStringSet(fasta, file=out_f, fo
```

②　①

R Console

```
> fasta <- getSeq(FaFile(in_f1), hoge)    #配列情報を取得した$
> fasta                                   #確認してるだけです
  A DNA ringSet instance of length 2262
        width seq                                names           $
    [1]  1317 GTGACTGATTTAGAA...AGCTAAAGCCATAG Chromosome
    [2]  1140 ATGAAATTTACAATT...TTAGAACTTACTAA Chromosome
    [3]   225 GTGCAAGAAGCAAAA...TTCAAAATGAGTAG Chromosome
    [4]  1122 ATGATTTTAAAAGAA...AGGAGGAACCATAG Chromosome
    [5]  1944 GTGAGCGATAAAAAA...ACTTAGATCTATAG Chromosome
    ...   ... ...
 [2258]  1389 GTGGCACAGACAGAG...GTTTAGGTAAATAG Chromosome
 [2259]   801 ATGGCAATTTTTACT...CTAGTGAGATGTAA Chromosome
 [2260]   834 GTGAAAAAGCACTTA...GTAGGCGCAAGTGA Chromosome
 [2261]   345 ATGAGAAAGTCATAT...TAGATGAGCATTAA Chromosome
 [2262]   135 ATGAAGCGCACATTT...TATTATCTGCATAG Chromosome
>
```

転写物配列取得

①のfastaオブジェクトをそのままFASTA形式で保存すると、②で見えているがままのdescription情報が書きだされる。つまり、すべて”Chromosome”になってしまう

```
#本番(配列取得)
fasta <- getSeq(FaFile(in_f1), hoge)    #配列情報を取得した結果をfastaに格納
fasta                                   #確認してるだけです

#後処理(description部分を変更)
names(fasta) <- paste(seqnames(hoge), start(ranges(hoge)),#"染色体名_start_end"に変更
                      end(ranges(hoge)), sep="_")#"染色体名_start_end"に変更
fasta                                   #確認してるだけです

#ファイルに保存
writeXStringSet(fasta, file=out_f, format="fasta", width=50)#fastaの中身を指定したファイル名で保存
```

①

R Console

```
> fasta <- getSeq(FaFile(in_f1), hoge)    #配列情報を取得した$
> fasta                                   #確認してるだけです
  A DNAStringSet instance of length 2262
        width seq                                         names               $
    [1]  1317 GTGACTGATTTAGAA...AGCTAAAGCCATAG            Chromosome
    [2]  1140 ATGAAATTTACAATT...TTAGAACTTACTAA            Chromosome
    [3]   225 GTGCAAGAAGCAAAA...TTCAAAATGAGTAG            Chromosome
    [4]  1122 ATGATTTTAAAAGAA...AGGAGGAACCATAG            Chromosome
    [5]  1944 GTGAGCGATAAAAAA...ACTTAGATCTATAG            Chromosome
    ...   ... ...
 [2258]  1389 GTGGCACAGACAGAG...GTTTAGGTAAATAG            Chromosome
 [2259]   801 ATGGCAATTTTTACT...CTAGTGAGATGTAA            Chromosome
 [2260]   834 GTGAAAAAGCACTTA...GTAGGCGCAAGTGA            Chromosome
 [2261]   345 ATGAGAAAGTCATAT...TAGATGAGCATTAA            Chromosome
 [2262]   135 ATGAAGCGCACATTT...TATTATCTGCATAG            Chromosome
> |
```

②

転写物配列取得

```
#本番(配列取得)
fasta <- getSeq(FaFile(in_f1), hoge)    #配列情報を取得し
fasta                                   #確認してるだけで

#後処理(description部分を変更)
names(fasta) <- paste(seqnames(hoge), start(ranges(hoge)),#"染色体名
                      end(ranges(hoge)), sep="_")#"染色体名_start_en
fasta                                   #確認してるだけです

#ファイルに保存
writeXStringSet(fasta, file=out_f, format="fasta", width=50)#fastaの中
```

①

②

赤枠部分で行っているのは、description部分の記述内容を“Chromosome_start_end”として、どこの座標由来の塩基配列かがわかるようにしている。①pasteは、文字列を②sepオプションで指定した文字を間に挟んで連結する関数。③の例をみれば挙動がわかると期待

③

R Console

```
> paste("uge", "age", sep="_")
[1] "uge_age"
> seqnames(hoge)
factor-Rle of length 2262 with 1 run
  Lengths:       2262
  Values : Chromosome
Levels(1): Chromosome
> ranges(hoge)
IRanges of length 2262
          start     end width
[1]         360    1676  1317
[2]        1852    2991  1140
[3]        3233    3457   225
[4]        3467    4588  1122
[5]        4588    6531  1944
...         ...     ...   ...
[2258] 2273924 2275312  1389
[2259] 2275488 2276288   801
[2260] 2276455 2277288   834
[2261] 2277304 2277648   345
[2262] 2277719 2277853   135
>
```

転写物配列取得

①description部分が変わっていることがわかる。これを眺めるだけで、出力ファイルをみなくてもうまくいっていると判断できる（と油断していると時々落とし穴があるので注意）

```
#本番(配列取得)
fasta <- getSeq(FaFile(in_f1), hoge)    #配列情報を取得した結果をfastaに格納
fasta                                   #確認してるだけです

#後処理(description部分を変更)
names(fasta) <- paste(seqnames(hoge), start(ranges(hoge)),#"染色体名_start_end"に変更
                      end(ranges(hoge)), sep="_")#"染色体名_start_end"に変更
fasta                                   #確認してるだけです

#ファイルに保存
writeXStringSet(fasta, file=out_f, forma
```

R Console

```
> #後処理(description部分を変更)
> names(fasta) <- paste(seqnames(hoge), start(ranges(hoge))$
+                       end(ranges(hoge)), sep="_")#"染色$
> fasta                                   #確認してるだけで$
  A DNAStringSet instance of length 2262
       width seq                          names
   [1]  1317 GTGACTGATTT...AAAGCCATAG Chromosome_360_1676
   [2]  1140 ATGAAATTTAC...AACTTACTAA Chromosome_1852_2991
   [3]   225 GTGCAAGAAGC...AAATGAGTAG Chromosome_3233_3457
   [4]  1122 ATGATTTTAAA...GGAACCATAG Chromosome_3467_4588
   [5]  1944 GTGAGCGATAA...AGATCTATAG Chromosome_4588_6531
    ...   ... ...
[2258]  1389 GTGGCACAGAC...AGGTAAATAG Chromosome_227392...
[2259]   801 ATGGCAATTTT...TGAGATGTAA Chromosome_227548...
[2260]   834 GTGAAAAAGCA...GCGCAAGTGA Chromosome_227645...
[2261]   345 ATGAGAAAGTC...TGAGCATTAA Chromosome_227730...
[2262]   135 ATGAAGCGCAC...ATCTGCATAG Chromosome_227771...
>
```

①

課題2

①このfastaオブジェクトを入力として、転写物数、塩基長の最大(max)・最小(min)・平均(mean)を示せ

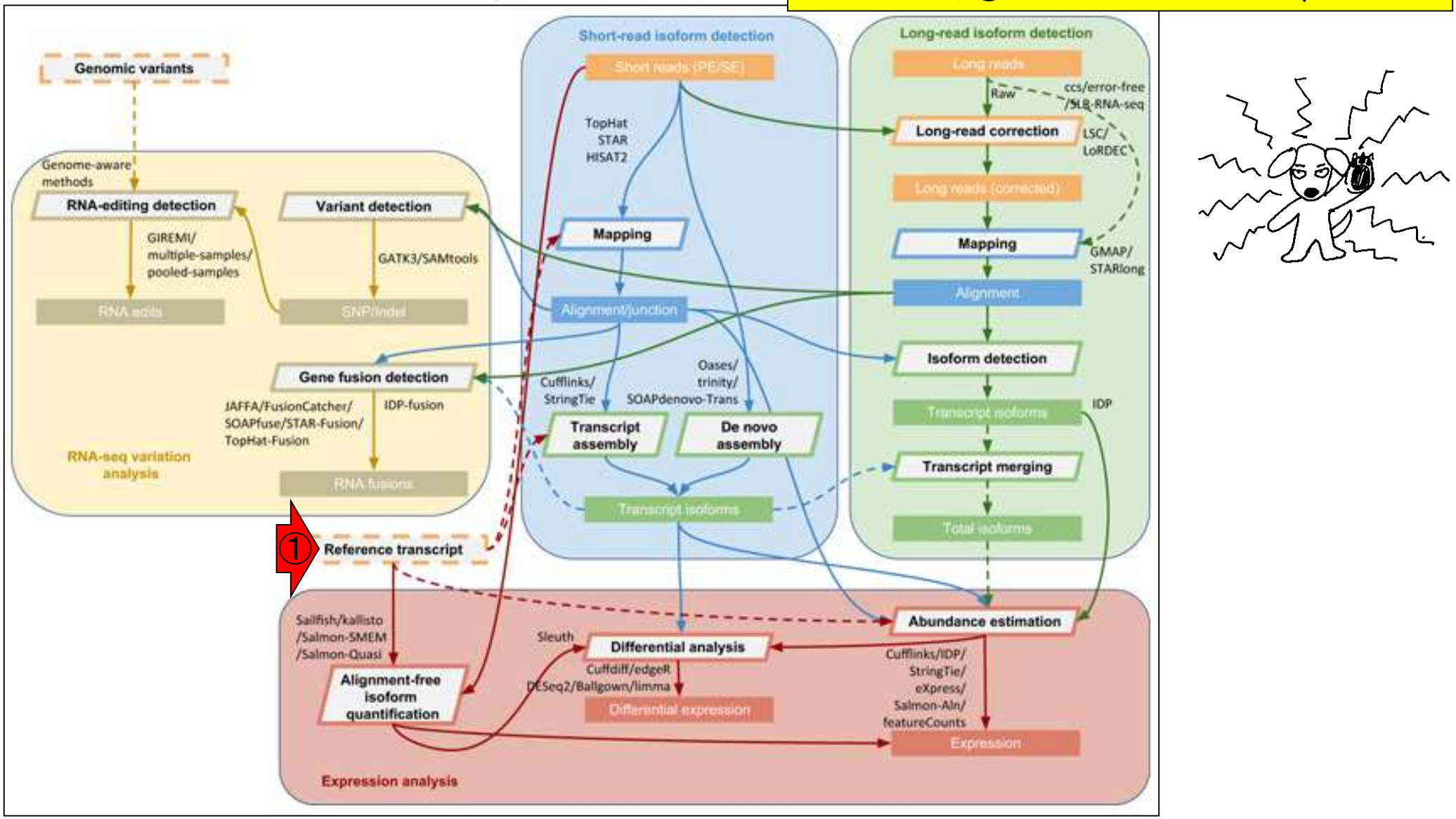
```
#本番(配列取得)
fasta <- getSeq(FaFile(in_f1), hoge)    #配列情報を取得した結果をfastaに格納
fasta                                   #確認してるだけです

#後処理(description部分を変更)
names(fasta) <- paste(seqnames(hoge), start(ranges(hoge)),#"染色体名_start_end"に変更
                      end(ranges(hoge)), sep="_")#"染色体名_start_end"に変更
fasta                                   #確認してるだけです

#ファイルに保存
writeXStringSet(fasta, file=out_f, forma
```

R Console

```
> #後処理(description部分を変更)
> names(fasta) <- paste(seqnames(hoge), start(ranges(hoge))$
+                       end(ranges(hoge)), sep="_")#"染色$
> fasta                                    #確認してるだけで$
  A DNAStringSet instance of length 2262
       width seq                          names
   [1]  1317 GTGACTGATTT...AAAGCCATAG Chromosome_360_1676
   [2]  1140 ATGAAATTTAC...AACTTACTAA Chromosome_1852_2991
   [3]   225 GTGCAAGAAGC...AAATGAGTAG Chromosome_3233_3457
   [4]  1122 ATGATTTTAAA...GGAACCATAG Chromosome_3467_4588
   [5]  1944 GTGAGCGATAA...AGATCTATAG Chromosome_4588_6531
    ...   ... ...
[2258]  1389 GTGGCACAGAC...AGGTAAATAG Chromosome_227392...
[2259]   801 ATGGCAATTTT...TGAGATGTAA Chromosome_227548...
[2260]   834 GTGAAAAAGCA...GCGCAAGTGA Chromosome_227645...
[2261]   345 ATGAGAAAGTC...TGAGCATTAA Chromosome_227730...
[2262]   135 ATGAAGCGCAC...ATCTGCATAG Chromosome_227771...
>
```

RNACocktail論文の図

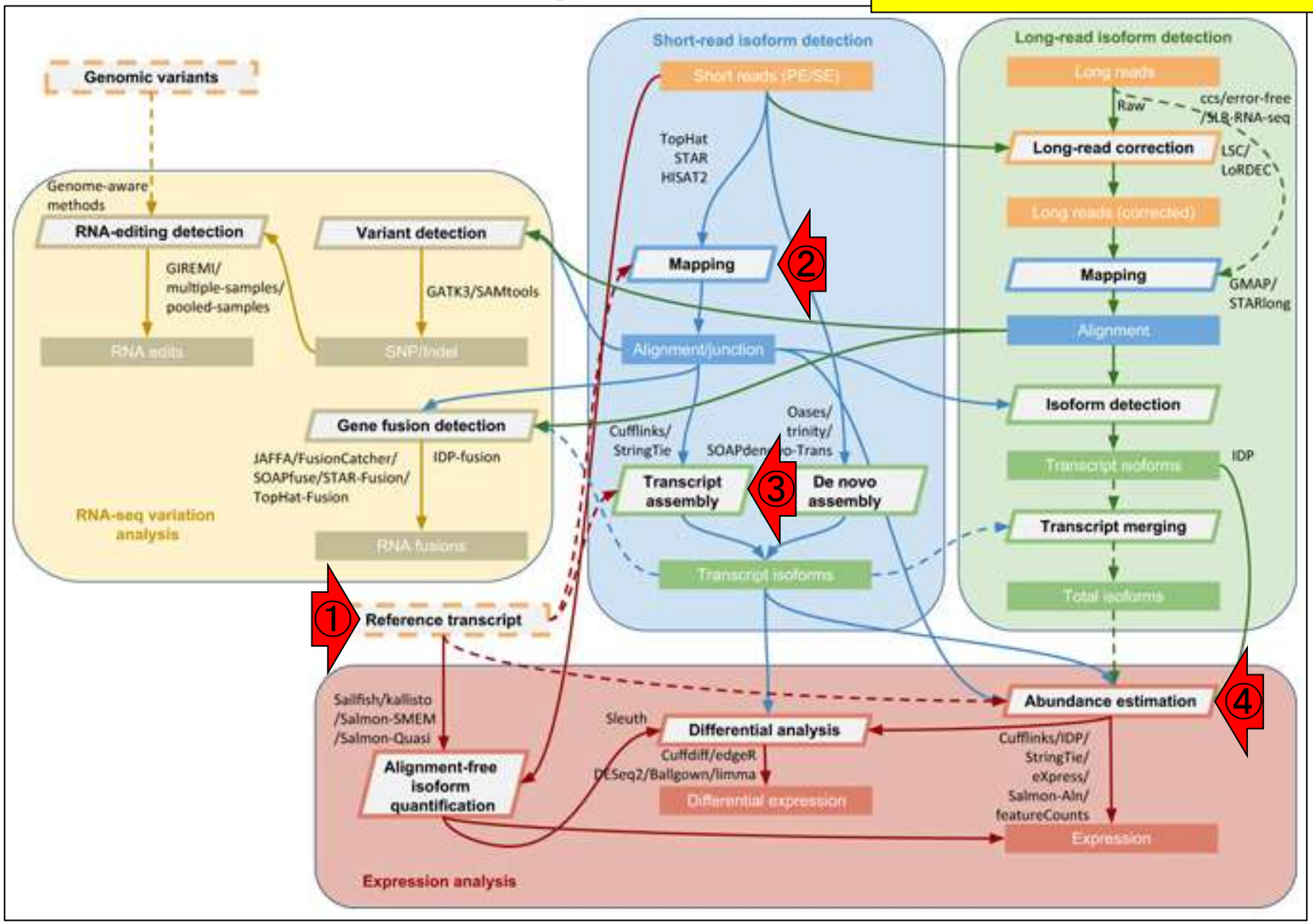
課題の位置づけについて説明。課題2で作成するトランスクリプトーム配列のmulti-FASTAファイルが、①のReference transcriptです。



Sahraeian et al., Nat Commun., **8**(1): 59, 2017

RNACocktail論文の図

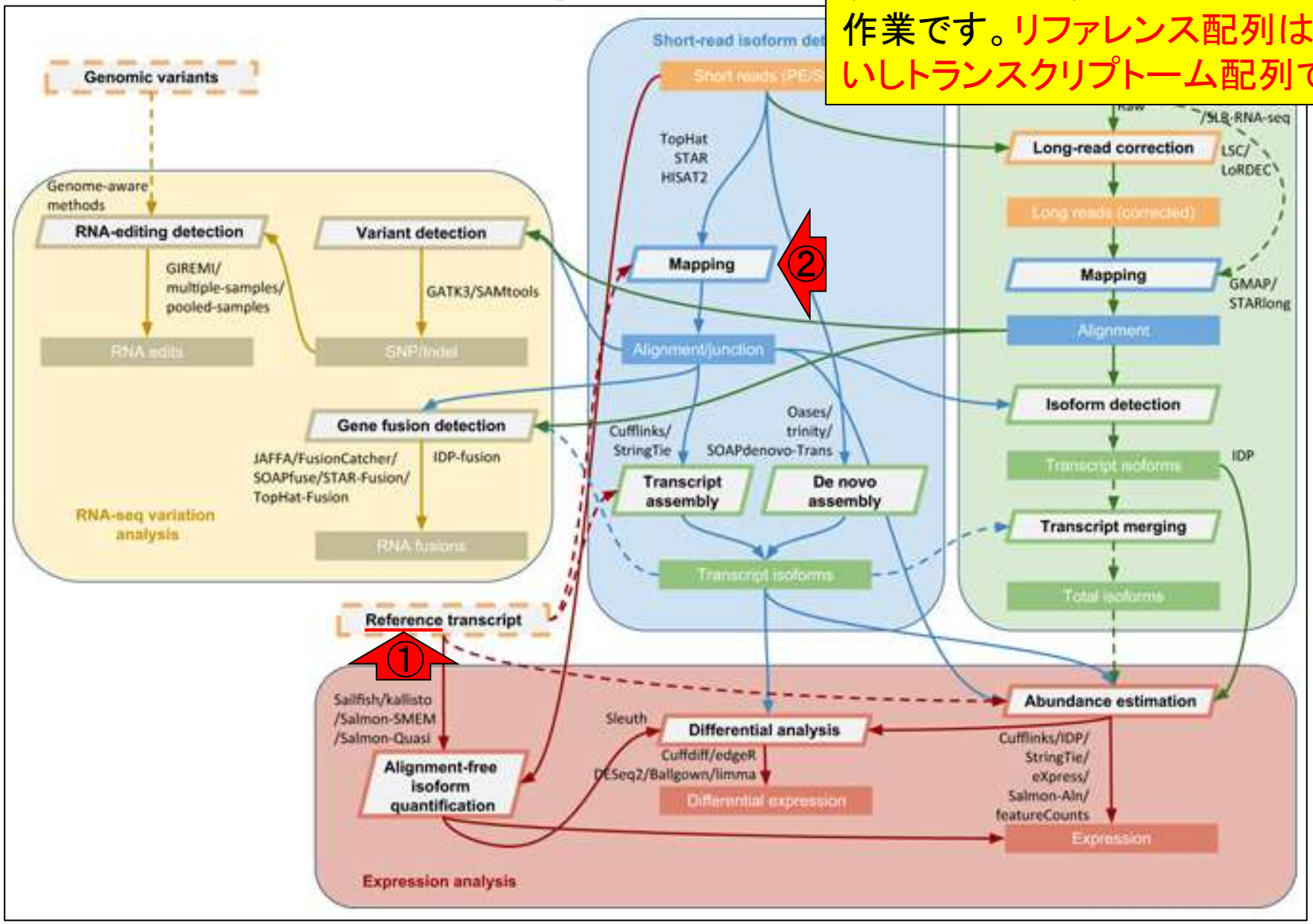
①のReference transcriptは、②Mapping、③Transcript assembly、④Abundance estimationなどでリファレンス配列として使われます



Sahraeian et al., Nat Commun., 8(1): 59, 2017

RNACocktail論文の図

①「リファレンス」は「参照」という意味。②マッピングは、RNA-seqリードをマップする側として用い、**リファレンス配列**のどこにマップされるかを調べる作業です。リファレンス配列は、ゲノム配列でもよいしトランスクリプトーム配列でもよいのです

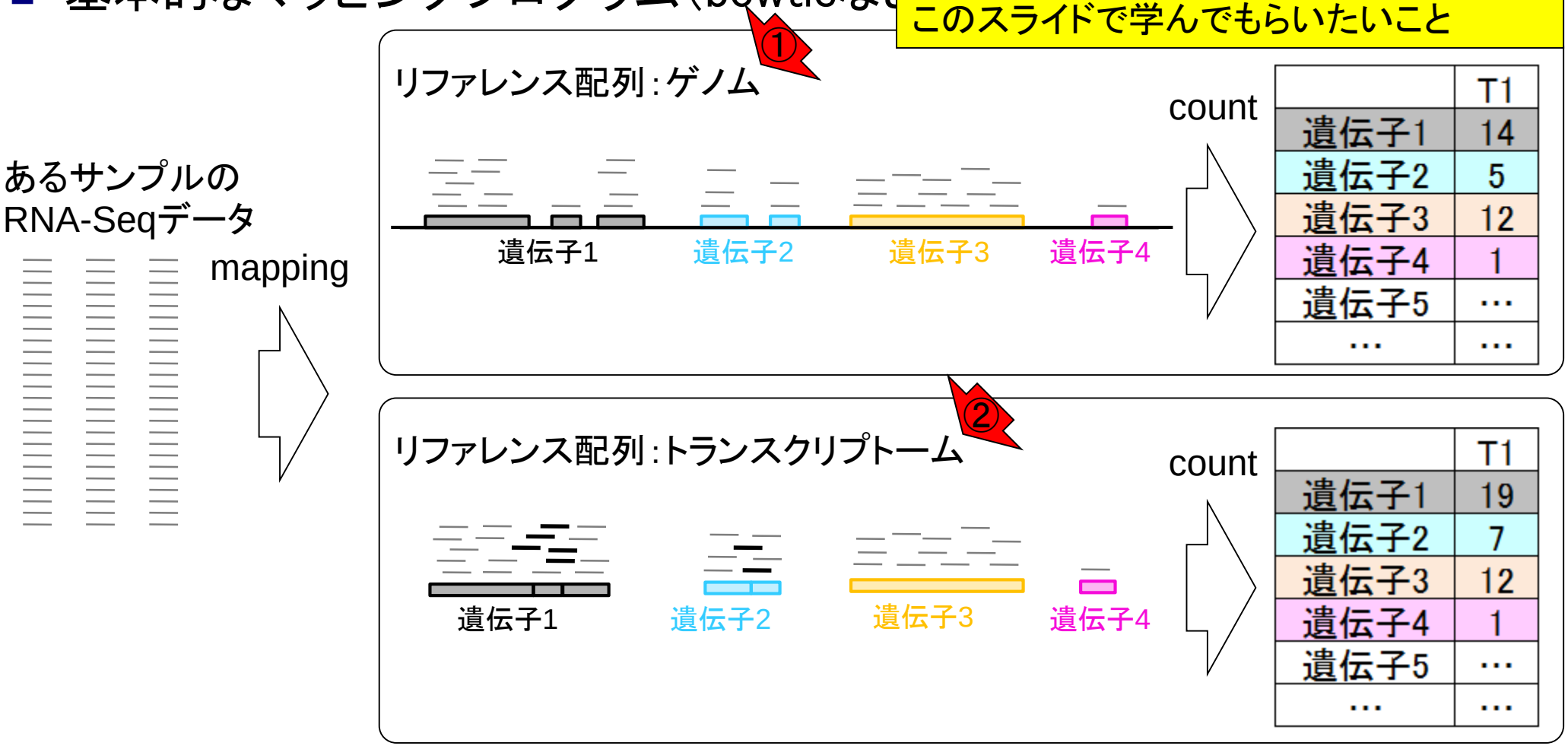


Sahraeian et al., Nat Commun., **8**(1): 59, 2017

マッピングのイメージ

- 基本的なマッピングプログラム（bowtieなど）

（遺伝子＝転写物ではないので若干不正確ではあるが…）①ゲノム配列以外に②トランスクリプトーム配列もリファレンスとして使える、という感覚を掴んでもらうのがこのスライドで学んでもらいたいこと



マッピングのイメージ

- 基本的なマッピングプログラム（bo

①トランスクリプトーム配列をリファレンスとして使うメリットは、リファレンスのサイズ（トータルの配列長と同義）がゲノムに比べて圧倒的に小さいので、マッピングがサクッと終わります。また、②で示したようなexon間にまたがるジャンクションリードのマッピングも、リファレンスがトランスクリプトーム配列の場合は、全く気にする必要がありません

あるサンプルの
RNA-Seqデータ

mapping

リファレンス配列：ゲノム

遺伝子1　遺伝子2　遺伝子3　遺伝子4

遺伝子2	5
遺伝子3	12
遺伝子4	1
遺伝子5	…
…	…

①

リファレンス配列：トランスクリプトーム

②

遺伝子1　遺伝子2　遺伝子3　遺伝子4

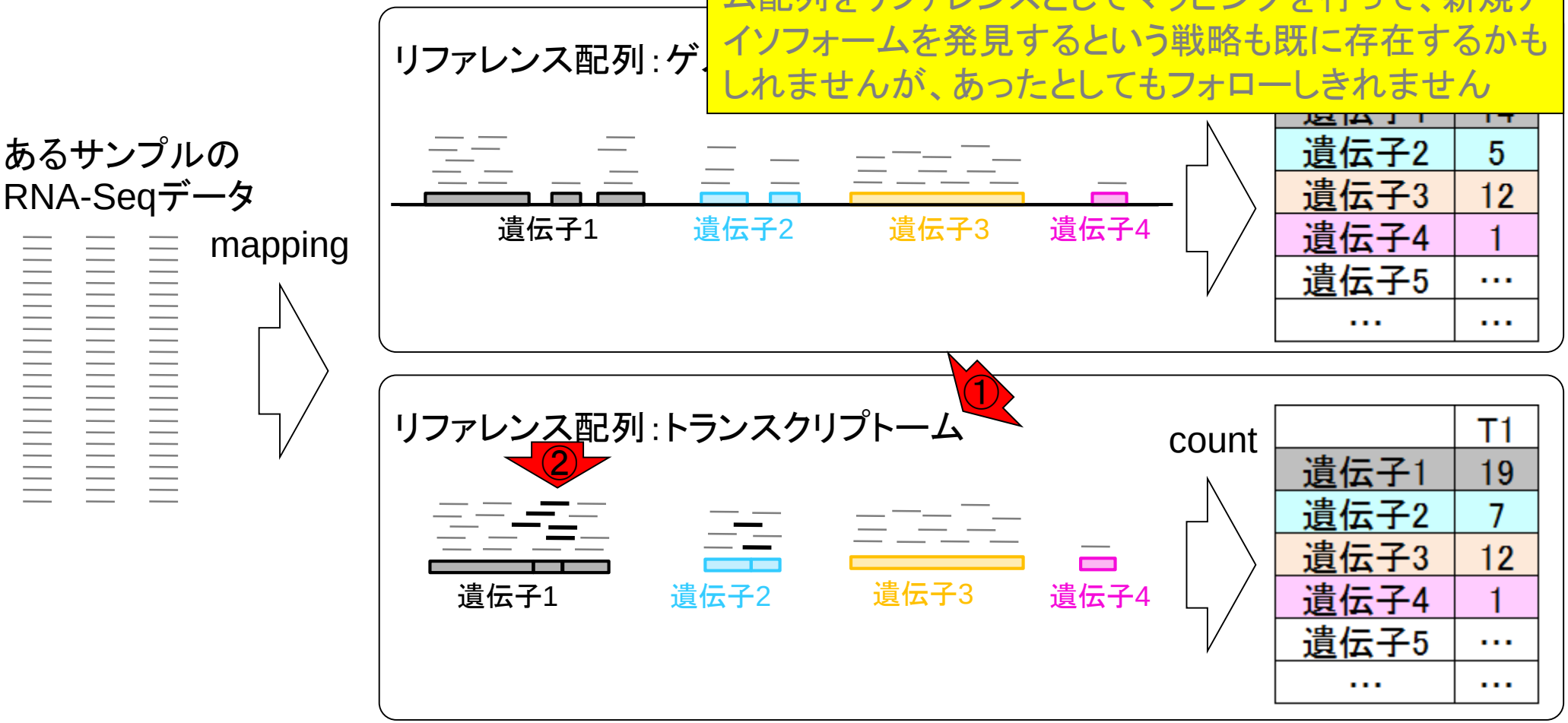
count

	T1
遺伝子1	19
遺伝子2	7
遺伝子3	12
遺伝子4	1
遺伝子5	…
…	…

マッピングのイメー

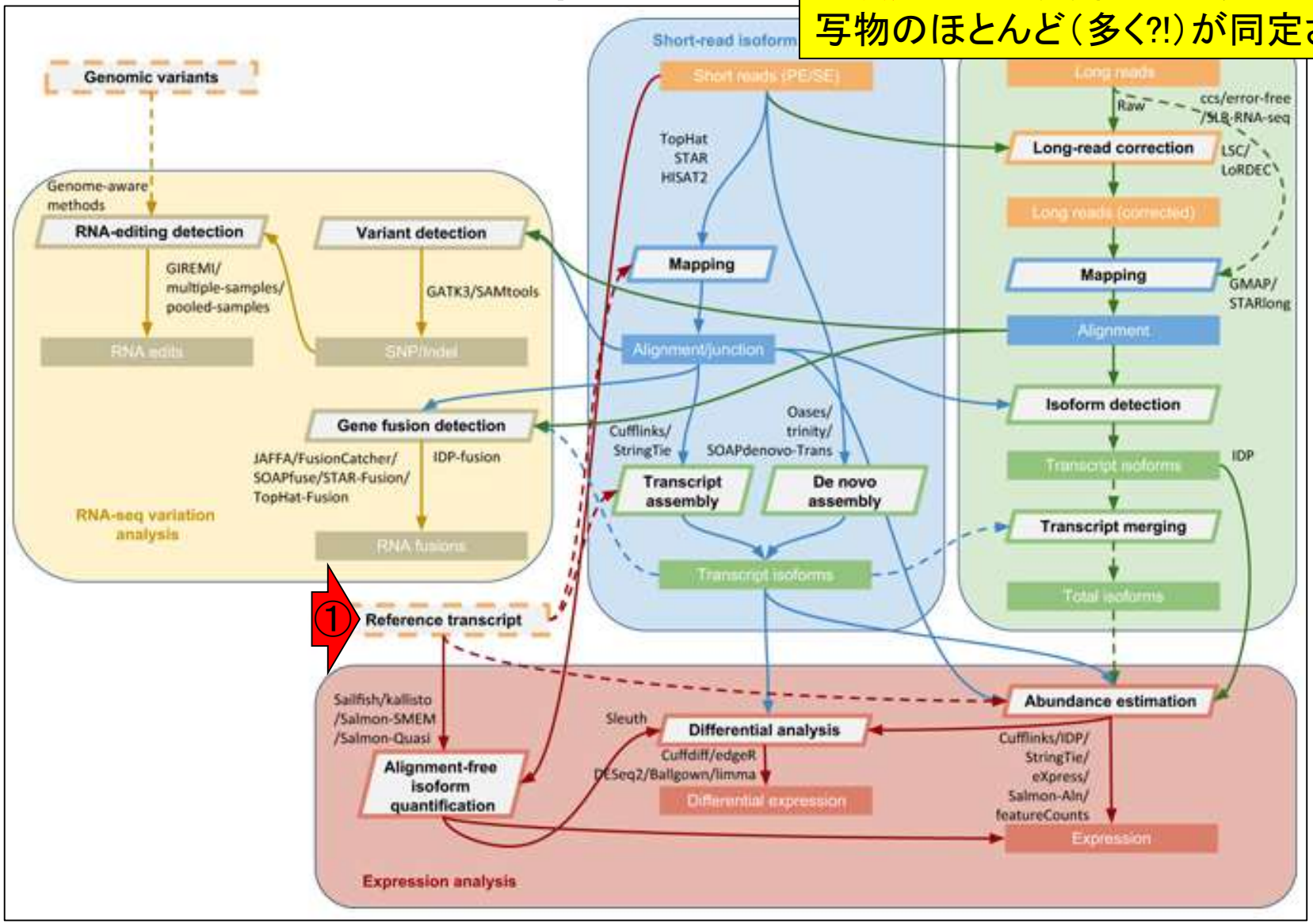
- 基本的なマッピングプログラム(

デメリットは、「トランスクリプトーム配列はこれだ！」と決め打ちしているようなものなので、新規転写物探しには向きません。それが目的の場合は、通常はゲノム配列をリファレンスとして用います。トランスクリプトーム配列をリファレンスとしてマッピングを行って、新規アイソフォームを発見するという戦略も既に存在するかもしれませんが、あったとしてもフォローしきれません



RNACocktail論文の

それゆえ、①リファレンスとしてトランスクリプトーム配列を用いるのは、実質的にヒトやマウスの場合に限定されます。様々な臓器や組織で発現する転写物のほとんど（多く?!）が同定されているからです



Sahraeian et al., Nat Commun., 8(1): 59, 2017

Contents

- トランスクリプトーム解析技術の原理や特徴
  - RNA-seq（Illuminaの場合）、遺伝子≠転写物
- RNA-seqデータ解析のイメージ
  - マッピング→新規転写物の同定
- 様々な解析目的
  - 転写物配列取得、手法比較論文の紹介、ウェブツール
  - データ解析の全体像（入出力の関係や代表的なツール）
- アノテーションファイルの読み込みと課題1
  - Rで転写物配列取得のイントロ
- Rで転写物配列取得と課題2
  - アノテーションファイルとゲノム情報ファイルから
- 公共データベース
  - NGS全体（NCBI SRA, EMBL-EBI ENA, DDBJ SRA）
  - DRAの概要、クオリティスコアなど

公共DB

①の公共DBからで示しているのは、トランスクリプトームに限らずNGS全体の話。②FASTQ形式ファイルは、データ解析の事実上の出発点

- イントロ | 一般 | 配列取得 | トランスクリプトーム配列 | GenomicFeatures(Lawrence 2013)(last modified 2(
- イントロ | 一般 | 配列取得 | トランスクリプトーム配列 | biomaRt(Durinck 2009)(last modified 2015/02/20)
- イントロ | 一般 | 読み込み | xlsx形式 | openxlsx(last modified 2015/11/15)
- イントロ | NGS | 様々なプラットフォーム(last modified 2016/03/24)
- イントロ | NGS | qPCRやmicroarrayなどとの比較 (last modified 2014/11/12)
- イントロ | NGS | 可視化(ゲノムブラウザやViewer) (last mod d 2016/12/22)
- イントロ | NGS | 配列取得 | FASTQ or SRA | 公共DBから ① modified 2015/02/23)
- イントロ | NGS | 配列取得 | FASTQ or SRA | SRAdb(Zhu 2(5) (last modified 2015/02/24)
- イントロ | NGS | 配列取得 | シミュレーションデータ | について (last modified 2015/01/18)
- イントロ | NGS | 配列取得 | シミュレーションデータ | ランダムな塩基配列の生成から (last modified 2015/01/18)
- イントロ | NGS | アノテーション情報取得 | について (last modified 2014/03/26)
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取

②

イントロ | NGS | 配列取得 | FASTQ or SRA | 公共DBから NEW

次世代シーケンサ(NGS)から得られる塩基配列データを公共データベースから取得する際には以下を利用します。マイクロアレイデータ取得のときと同様、NGSデータもArrayExpress経由でダウンロードするのがいいかもしれません。メタデータの全貌を把握しやすいこと、生データ(raw data)だけでなく加工済みのデータ(processed data)がある場合にはその存在がすぐにわかることなど、操作性の点で他を凌駕していると思います。上記でも触れているようにFASTQファイルのダウンロードからマッピングまでを行うのはエンドユーザレベルでは大変ですが、submitterが提供してくれている場合は(まだまだ少ないようですが)リファレンス配列へのマップ後のデータ、つまりBAM形式ファイルの提供もすでに始まっているようです。2014年6月26日に知りました(DDBJ児玉さんありがとうございますm(_ _)m)。

データの形式は基本的にSanger typeのFASTQ形式です。FASTA形式はリードあたり二行(idの行と配列の行)で表現します。FASTQ形式はリードあたり4行(@から始まるidの行と配列の行、および+から始まるidの行とbase callの際のqualityの行)で表現します。FASTQ形式は、Sangerのものがデファクトスタンダード(業界標準)です。かつてIlluminaのプラットフォームから得られるのはFASTQ-like formatという表現がなされていたようです(Cock et al., Nucleic Acids Res., 2010)。しかし少なくとも2013年頃には、IlluminaデータもBaseSpaceやCASAVA1.8のconfigureBclToFastq.plなどを用いることで業界標準のFASTQ形式(つまりSanger typeのデータ)に切り替えられるようですし、NCBI SRAなどの公共DBから取得するデータは全てはSanger typeのデータになっていたと思います(Kibukawa E., テクニカルサポートウェビナー, 2013)。

- DDBJ Sequence Read Archive (DRA): Kodama et al., Nucleic Acids Res., 2012
- EMBL-EBI European Nucleotide Archive (ENA): Silvester et al., Nucleic Acids Res., 2015
- NCBI Sequence Read Archive (SRA): NCBI Resource Coordinators., Nucleic Acids Res., 2014
- ArrayExpress: Kolesnikov et al., Nucleic Acids Res., 2015
- GEO: Barrett et al., Nucleic Acids Res., 2013
- DBCLS SRA: Nakazato et al., PLoS One, 2013

公共DB

NGSデータの公共DBは、①日本(のDDBJという組織)、②米国(NCBI)、そして欧州(EMBL-EBI)の三極で運用されている

- イントロ | 一般 | 配列取得 | トランスクリプトーム配列 | GenomicFeatures(Lawrence 2013)(last modified 2016/02/10)
- イントロ | 一般 | 配列取得 | トランスクリプトーム配列 | biomaRt(Durinck 2009)(last modified 2015/02/20)
- イントロ | 一般 | 読み込み | xlsx形式 | openxlsx(last modified 2015/11/15)
- イントロ | NGS | 様々なプラットフォーム(last modified 2016/03/24)
- イントロ | NGS | qPCRやmicroarrayなどとの比較 (last modified 2014/11/12)
- イントロ | NGS | 可視化(ゲノムブラウザやViewer) (last modified 2016/12/22)
- イントロ | NGS | 配列取得 | FASTQ or SRA | 公共DBから (last modified 2015/02/23)
- イントロ | NGS | 配列取得 | FASTQ or SRA | SRAdb(Zhu 2013) (last modified 2015/02/24)
- イントロ | NGS | 配列取得 | シミュレーションデータ | について (last modified 2015/01/18)
- イントロ | NGS | 配列取得 | シミュレーションデータ | ランダムな塩基配列の生成から (last modified 2015/01/18)
- イントロ | NGS | アノテーション情報取得 | について (last modified 2014/03/26)
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取

イントロ | NGS | 配列取得 | FASTQ or SRA | 公共DBから NEW

次世代シーケンサ(NGS)から得られる塩基配列データを公共データベースから取得する際には以下を利用します。マイクロアレイデータ取得のときと同様、NGSデータもArrayExpress経由でダウンロードするのがいいかもしれません。メタデータの全貌を把握しやすいこと、生データ(raw data)だけでなく加工済みのデータ(processed data)がある場合にはその存在がすぐにわかることなど、操作性の点で他を凌駕していると思います。上記でも触れているようにFASTQファイルのダウンロードからマッピングまでを行うのはエンドユーザレベルでは大変ですが、submitterが提供してくれている場合は(まだまだ少ないようですが)リファレンス配列へのマップ後のデータ、つまりBAM形式ファイルの提供もすでに始まっているようです。2014年6月26日に知りました(DDBJ児玉さんありがとうございますm(_ _)m)。

データの形式は基本的にSanger typeのFASTQ形式です。FASTA形式はリードあたり二行(idの行と配列の行)で表現します。FASTQ形式はリードあたり4行(@から始まるidの行と配列の行、および+から始まるidの行とbase callの際のqualityの行)で表現します。FASTQ形式は、Sangerのものがデファクトスタンダード(業界標準)です。かつてIlluminaのプラットフォームから得られるのはFASTQ-like formatという表現がなされていたようです(Cock et al., Nucleic Acids Res., 2010)。しかし少なくとも2013年頃には、IlluminaデータもBaseSpaceやCASAVA1.8のconfigureBclToFastq.plなどを用いることで業界標準のFASTQ形式(つまりSanger typeのデータ)に切り替えられるようですし、NCBI SRAなどの公共DBから取得するデータは全てはSanger typeのデータになっていたと思います(Kibukawa E., テクニカルサポートウェビナー, 2013)。

① DDBJ Sequence Read Archive (DRA): Kodama et al., Nucleic Acids Res., 2012
② EMBL-EBI European Nucleotide Archive (ENA): Silvester et al., Nucleic Acids Res., 2015
③ NCBI Sequence Read Archive (SRA): NCBI Resource Coordinators., Nucleic Acids Res., 2014
ArrayExpress: Kolesnikov et al., Nucleic Acids Res., 2015
- GEO: Barrett et al., Nucleic Acids Res., 2013
- DBCLS SRA: Nakazato et al., PLoS One, 2013

大元はSRA形式ファイル

公共DBにある生データの大元は、①SRAと呼ばれる形式のファイル（拡張子が.sra）。②日、③米、④欧の三極ともに.sraをダウンロード可能

①

イントロ | NGS | 配列取得 | FASTQ or SRA | 公共DBから NEW

次世代シーケンサ(NGS)から得られる塩基配列データを公共データベースから取得する際には以下を利用します。マイクロアレイデータ取得のときと同様、NGSデータもArrayExpress経由でダウンロードするのがいいかもしれません。メタデータの全貌を把握しやすいこと、生データ(raw data)だけでなく加工済みのデータ(processed data)がある場合にはその存在がすぐにわかることなど、操作性の点で他を凌駕していると思います。上記でも触れているようにFASTQファイルのダウンロードからマッピングまでを行うのはエンドユーザレベルでは大変ですが、submitterが提供してくれている場合は（まだまだ少ないようですが）リファレンス配列へのマップ後のデータ、つまりBAM形式ファイルの提供もすでに始まっているようです。2014年6月26日に知りました(DDBJ児玉さんありがとうございますm(_ _)m)。

データの形式は基本的にSanger typeのFASTQ形式です。FASTA形式はリードあたり二行（idの行と配列の行）で表現します。FASTQ形式はリードあたり4行（@から始まるidの行と配列の行、および+から始まるidの行とbase callの際のqualityの行）で表現します。FASTQ形式は、Sangerのものがデファクトスタンダード（業界標準）です。かつてIlluminaのプラットフォームから得られるのはFASTQ-like formatという表現がなされていたようです(Cock et al., Nucleic Acids Res., 2010)。しかし少なくとも2013年頃には、IlluminaデータもBaseSpaceやCASAVA1.8のconfigureBclToFastq.plなどを用いることで業界標準のFASTQ形式(つまりSanger typeのデータ)に切り替えられるようですし、NCBI SRAなどの公共DBから取得するデータは全てはSanger typeのデータになっていたと思います(Kibukawa E., テクニカルサポートウェビナー, 2013)。

②
- DDBJ Sequence Read Archive (DRA): Kodama et al., Nucleic Acids Res., 2012
- EMBL-EBI European Nucleotide Archive (ENA): Silvester et al., Nucleic Acids Res., 2015

③
- NCBI Sequence Read Archive (SRA): NCBI Resource Coordinators., Nucleic Acids Res., 2014
- ArrayExpress: Kolesnikov et al., Nucleic Acids Res., 2015
  - GEO: Barrett et al., Nucleic Acids Res., 2013
  - DBCLS SRA: Nakazato et al., PLoS One, 2013

④

FASTQファイルは…

①FASTQ形式ファイル（拡張子が.fastqまたは.fq）を提供しているのは、②日（DRA）と③欧（ENA）のみ

- イントロ | 一般 | 配列取得 | トランスクリプトーム配列 | GenomicFeatures(Lawrence 2013)(last modified 2016/02/10)
- イントロ | 一般 | 配列取得 | トランスクリプトーム配列 | biomaRt(Durinck 2009)(last modified 2015/02/20)
- イントロ | 一般 | 読み込み | xlsx形式 | openxlsx(last modified 2015/11/15)
- イントロ | NGS | 様々なプラットフォーム(last modified 2016/03/24)
- イントロ | NGS | qPCRやmicroarrayなどとの比較 (last modified 2014/11/12)
- イントロ | NGS | 可視化(ゲノムブラウザやViewer) (last modified 2016/12/22)
- イントロ | NGS | 配列取得 | FASTQ or SRA | 公共DBから (last modified 2015/02/23)
- イントロ | NGS | 配列取得 | FASTQ or SRA | SRAdb(Zhu 2013) (last modified 2015/02/24)
- イントロ | NGS | 配列取得 | シミュレーションデータ | について (last modified 2015/01/18)
- イントロ | NGS | 配列取得 | シミュレーションデータ | ランダムな塩基配列の生成から (last modified 2015/01/18)
- イントロ | NGS | アノテーション情報取得 | について (last modified 2014/03/26)
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取

①

イントロ | NGS | 配列取得 | FASTQ or SRA | 公共DBから NEW

次世代シーケンサ(NGS)から得られる塩基配列データを公共データベースから取得する際には以下を利用します。マイクロアレイデータ取得のときと同様、NGSデータもArrayExpress経由でダウンロードするのがいいかもしれません。メタデータの全貌を把握しやすいこと、生データ(raw data)だけでなく加工済みのデータ(processed data)がある場合にはその存在がすぐにわかることなど、操作性の点で他を凌駕していると思います。上記でも触れているようにFASTQファイルのダウンロードからマッピングまでを行うのはエンドユーザレベルでは大変ですが、submitterが提供してくれている場合は（まだまだ少ないようですが）リファレンス配列へのマップ後のデータ、つまりBAM形式ファイルの提供もすでに始まっているようです。2014年6月26日に知りました(DDBJ児玉さんありがとうございますm(_ _)m)。

データの形式は基本的にSanger typeのFASTQ形式です。FASTA形式はリードあたり二行（idの行と配列の行）で表現します。FASTQ形式はリードあたり4行（@から始まるidの行と配列の行、および+から始まるidの行とbase callの際のqualityの行）で表現します。FASTQ形式は、Sangerのものがデファクトスタンダード（業界標準）です。かつてIlluminaのプラットフォームから得られるのはFASTQ-like formatという表現がなされていたようです(Cock et al., Nucleic Acids Res., 2010)。しかし少なくとも2013年頃には、IlluminaデータもBaseSpaceやCASAVA1.8のconfigureBclToFastq.plなどを用いることで業界標準のFASTQ形式(つまりSanger typeのデータ)に切り替えられるようですし、NCBI SRAなどの公共DBから取得するデータは全てはSanger typeのデータになっていたと思います(Kibukawa E., テクニカルサポートウェビナー, 2013)。

②

- DDBJ Sequence Read Archive (DRA): Kodama et al., Nucleic Acids Res., 2012
- EMBL-EBI European Nucleotide Archive (ENA): Silvester et al., Nucleic Acids Res., 2015
- NCBI Sequence Read Archive (SRA): NCBI Resource Coordinators., Nucleic Acids Res., 2014
- ArrayExpress: Kolesnikov et al., Nucleic Acids Res., 2015
- GEO: Barrett et al., Nucleic Acids Res., 2013
- DBCLS SRA: Nakazato et al., PLoS One, 2013

③

.sraから.fastqを作成

①DRAと②ENAは、③大元のSRAファイルを入力として、(Linux上で使えるfastq-dumpというプログラムを実行して)④FASTQファイルを作成し、それを提供しています。それゆえ、SRAファイルは公開済みでも、FASTQファイルの公開が場合によっては数か月後になることもあります。DB側のディスク容量の関係で同期を一時的にストップさせることもあるようです。

- イントロ | 一般 | 配列取得 | トランスクリプトーム配列 | GenomicFeatures(Lawrenc
- イントロ | 一般 | 配列取得 | トランスクリプトーム配列 | biomaRt(Durinck 2009)(la
- イントロ | 一般 | 読み込み | xlsx形式 | openxlsx(last modified 2015/11/15)
- イントロ | NGS | 様々なプラットフォーム(last modified 2016/03/24)
- イントロ | NGS | qPCRやmicroarrayなどとの比較 (last modified 2014/11/12)
- イントロ | NGS | 可視化(ゲノムブラウザやViewer) (last modified 2016/12/22)
- イントロ | NGS | 配列取得 | FASTQ or SRA | 公共DBから (last modified 2015/02
- イントロ | NGS | 配列取得 | FASTQ or SRA | SRAdb(Zhu 2013) (last modified 20
- イントロ | NGS | 配列取得 | シミュレーションデータ | について (last modified 2015/01/18)
- イントロ | NGS | 配列取得 | シミュレーションデータ | ランダムな塩基配列の生成から (last modified 2015/01/18)
- イントロ | NGS | アノテーション情報取得 | について (last modified 2014/03/26)
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取

④ ③

イントロ | NGS | 配列取得 | FASTQ or SRA | 公共DBから NEW

次世代シーケンサ(NGS)から得られる塩基配列データを公共データベースから取得する際には以下を利用します。マイクロアレイデータ取得のときと同様、NGSデータもArrayExpress経由でダウンロードするのがいいかもしれません。メタデータの全貌を把握しやすいこと、生データ(raw data)だけでなく加工済みのデータ(processed data)がある場合にはその存在がすぐにわかることなど、操作性の点で他を凌駕していると思います。上記でも触れているようにFASTQファイルのダウンロードからマッピングまでを行うのはエンドユーザレベルでは大変ですが、submitterが提供してくれている場合は(まだまだ少ないようですが)リファレンス配列へのマップ後のデータ、つまりBAM形式ファイルの提供もすでに始まっているようです。2014年6月26日に知りました(DDBJ児玉さんありがとうございますm(_ _)m)。

データの形式は基本的にSanger typeのFASTQ形式です。FASTA形式はリードあたり二行(idの行と配列の行)で表現します。FASTQ形式はリードあたり4行(@から始まるidの行と配列の行、および+から始まるidの行とbase callの際のqualityの行)で表現します。FASTQ形式は、Sangerのものがデファクトスタンダード(業界標準)です。かつてIlluminaのプラットフォームから得られるのはFASTQ-like formatという表現がなされていたようです(Cock et al., Nucleic Acids Res., 2010)。しかし少なくとも2013年頃には、IlluminaデータもBaseSpaceやCASAVA1.8のconfigureBclToFastq.plなどを用いることで業界標準のFASTQ形式(つまりSanger typeのデータ)に切り替えられるようですし、NCBI SRAなどの公共DBから取得するデータは全てはSanger typeのデータになっていたと思います(Kibukawa E., テクニカルサポートウェビナー, 2013)。

① ②

- DDBJ Sequence Read Archive (DRA): Kodama et al., Nucleic Acids Res., 2012
- EMBL-EBI European Nucleotide Archive (ENA): Silvester et al., Nucleic Acids Res., 2015
- NCBI Sequence Read Archive (SRA): NCBI Resource Coordinators., Nucleic Acids Res., 2014
- ArrayExpress: Kolesnikov et al., Nucleic Acids Res., 2015
- GEO: Barrett et al., Nucleic Acids Res., 2013
- DBCLS SRA: Nakazato et al., PLoS One, 2013

Contents

- トランスクリプトーム解析技術の原理や特徴
  - RNA-seq（Illuminaの場合）、遺伝子≠転写物
- RNA-seqデータ解析のイメージ
  - マッピング→新規転写物の同定
- 様々な解析目的
  - 転写物配列取得、手法比較論文の紹介、ウェブツール
  - データ解析の全体像（入出力の関係や代表的なツール）
- アノテーションファイルの読み込みと課題1
  - Rで転写物配列取得のイントロ
- Rで転写物配列取得と課題2
  - アノテーションファイルとゲノム情報ファイルから
- 公共データベース
  - NGS全体（NCBI SRA, EMBL-EBI ENA, DDBJ SRA）
  - DRAの概要、クオリティスコアなど

DRAを概観

①DRAをちょっと眺める。ネットワークの調子が悪い場合は、アクセスできたつもりで講義スライドを眺めていこう

イントロ | NGS | 配列取得 | FASTQ or SRA | 公共DBから NEW

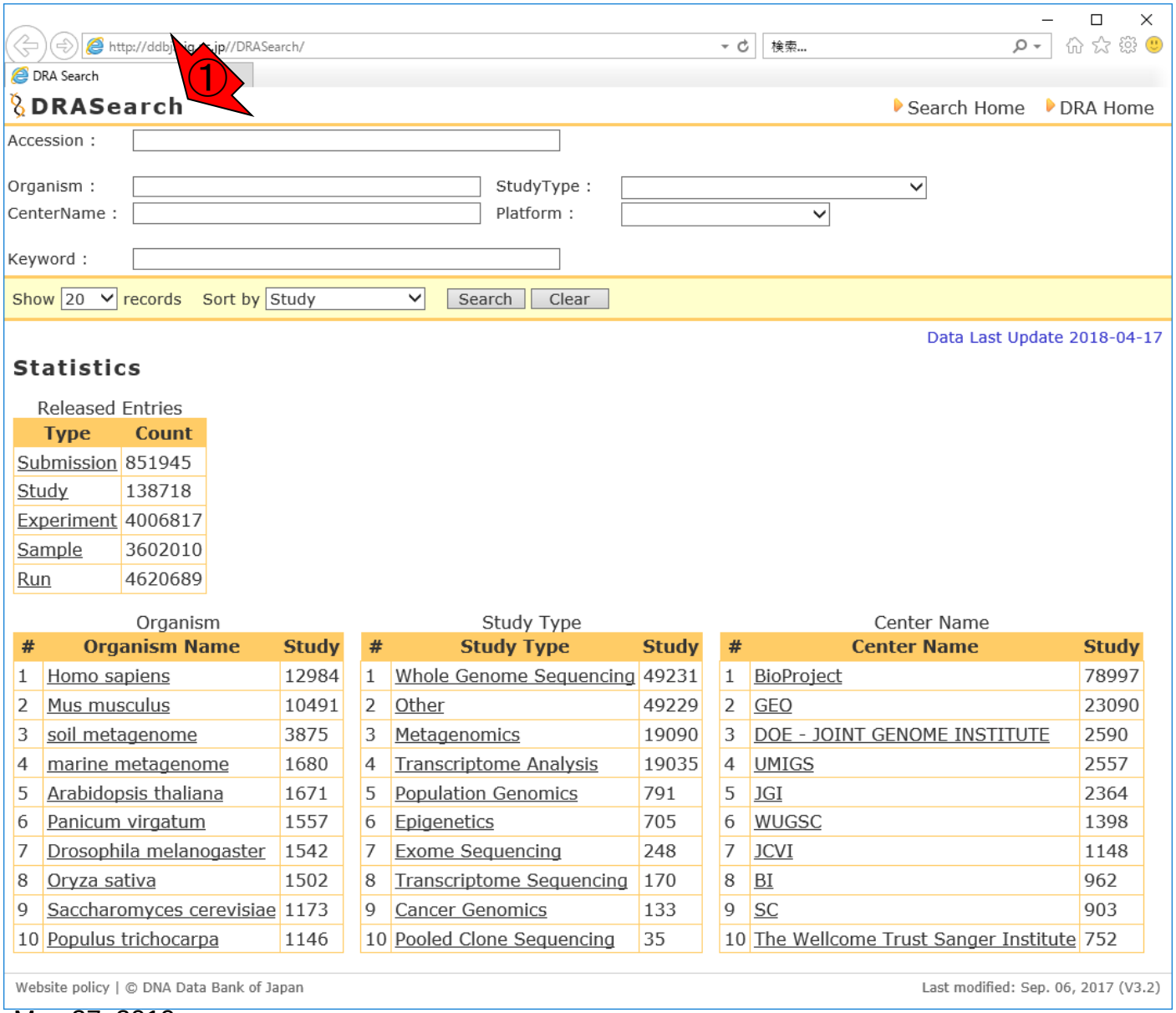
次世代シーケンサ(NGS)から得られる塩基配列データを公共データベースから取得する際には以下を利用します。マイクロアレイデータ取得のときと同様、NGSデータもArrayExpress経由でダウンロードするのがいいかもしれません。メタデータの全貌を把握しやすいこと、生データ(raw data)だけでなく加工済みのデータ(processed data)がある場合にはその存在がすぐにわかることなど、操作性の点で他を凌駕していると思います。上記でも触れているようにFASTQファイルのダウンロードからマッピングまでを行うのはエンドユーザレベルでは大変ですが、submitterが提供してくれている場合は(まだまだ少ないようですが)リファレンス配列へのマップ後のデータ、つまりBAM形式ファイルの提供もすでに始まっているようです。2014年6月26日に知りました(DDBJ児玉さんありがとうございますm(_ _)m)。

データの形式は基本的にSanger typeのFASTQ形式です。FASTA形式はリードあたり二行(idの行と配列の行)で表現します。FASTQ形式はリードあたり4行(@から始まるidの行と配列の行、および+から始まるidの行とbase callの際のqualityの行)で表現します。FASTQ形式は、Sangerのものがデファクトスタンダード(業界標準)です。かつてIlluminaのプラットフォームから得られるのはFASTQ-like formatという表現がなされていたようです(Cock et al., Nucleic Acids Res., 2010)。しかし少なくとも2013年頃には、IlluminaデータもBaseSpaceやCASAVA1.8のconfigureBclToFastq.plなどを用いることで業界標準のFASTQ形式(つまりSanger typeのデータ)に切り替えられるようですし、NCBI SRAなどの公共DBから取得するデータは全てはSanger typeのデータになっていたと思います(Kibukawa E., テクニカルサポートウェビナー, 2013)。

① DDBJ Sequence Read Archive (DRA): Kodama et al., Nucleic Acids Res., 2012
EMBL-EBI European Nucleotide Archive (ENA): Silvester et al., Nucleic Acids Res., 2015
- NCBI Sequence Read Archive (SRA): NCBI Resource Coordinators., Nucleic Acids Res., 2014
- ArrayExpress: Kolesnikov et al., Nucleic Acids Res., 2015
- GEO: Barrett et al., Nucleic Acids Res., 2013
- DBCLS SRA: Nakazato et al., PLoS One, 2013

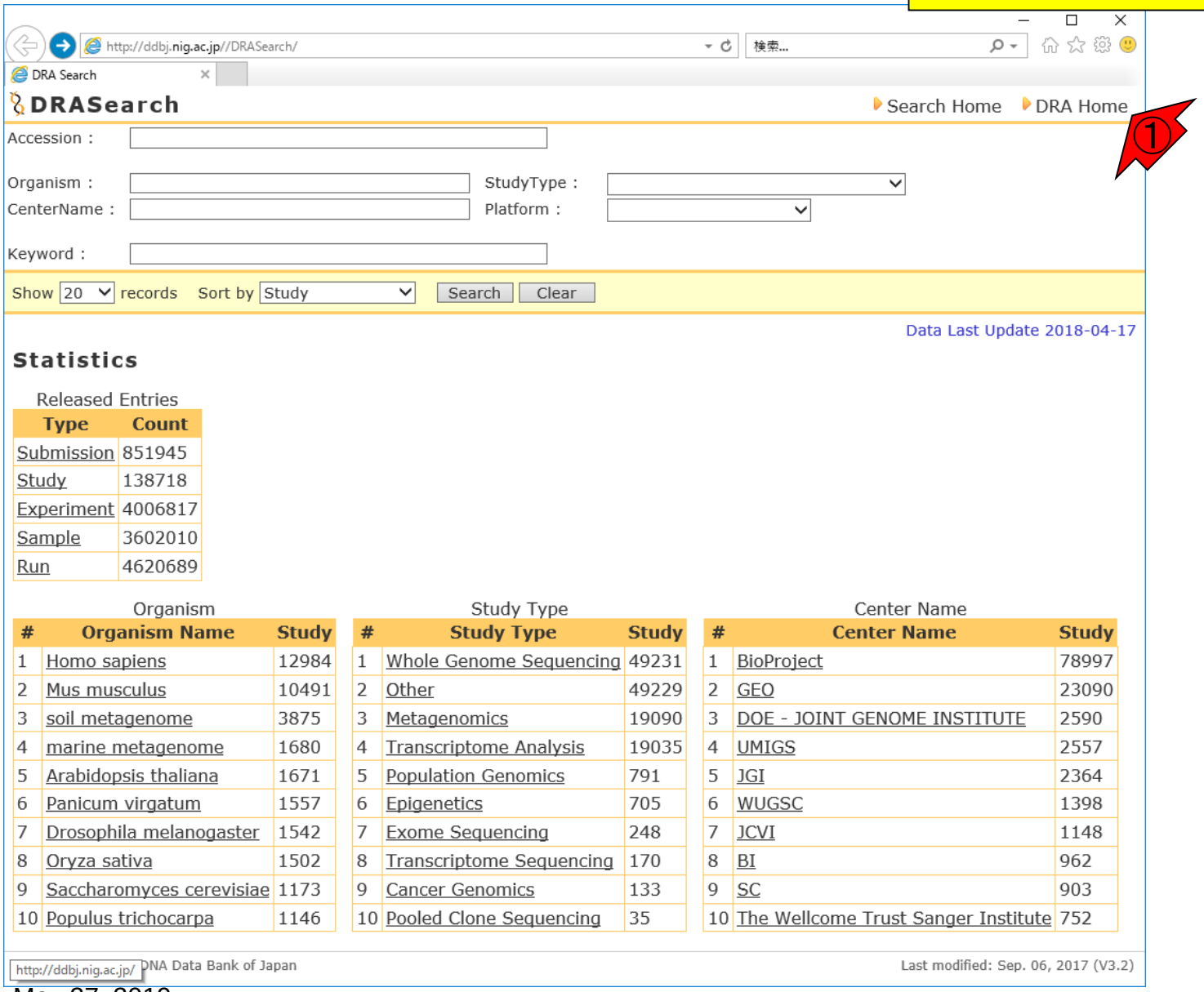
DRAを概観

①DRASearchというページに飛ばしています



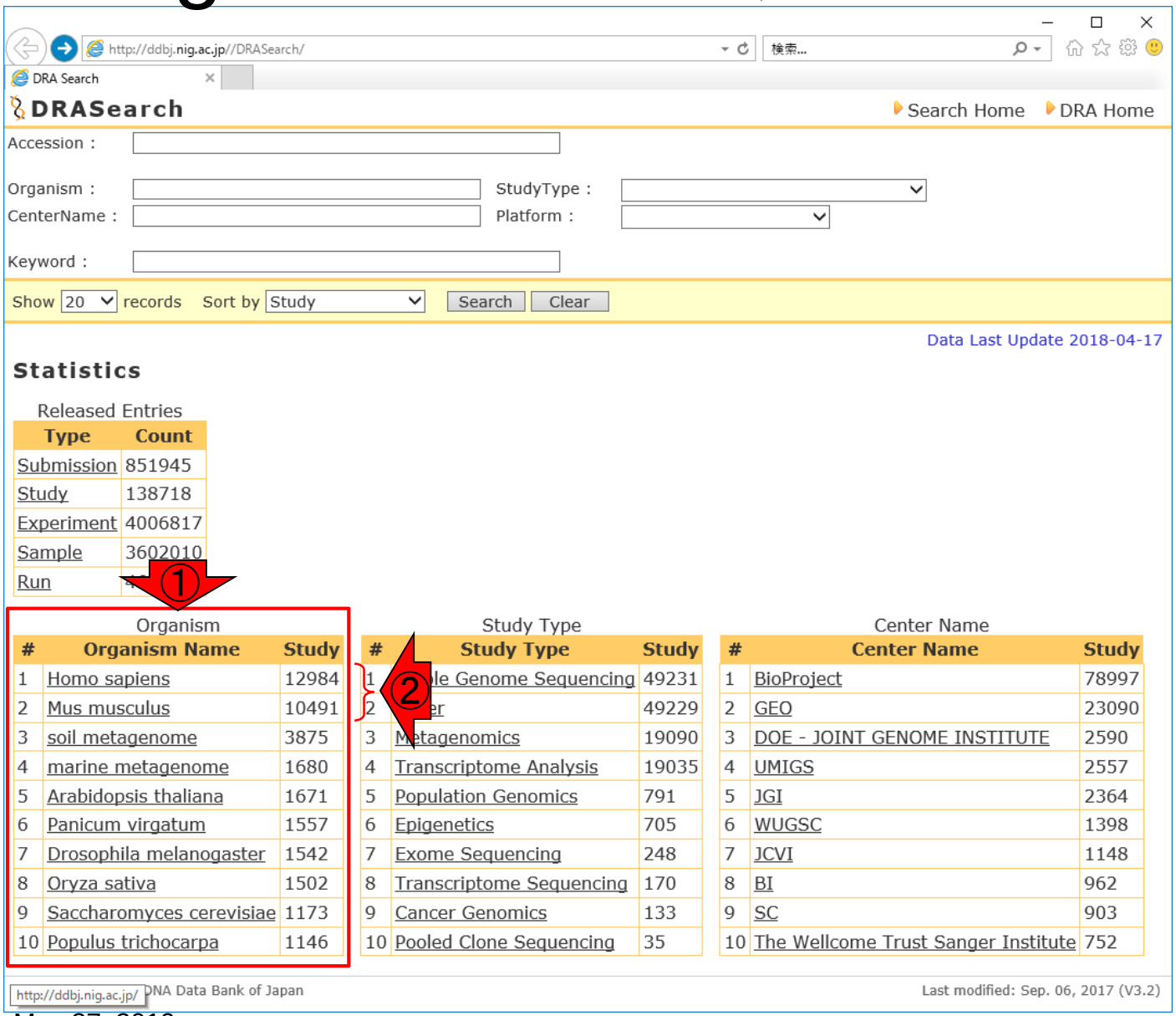
DRAを概観

①以前はDRA単体のトップページがありましたが、①のリンク先がDDBJ本体のトップページに飛ぶように最近?!切り替わったようです



Organismでの分類

①生物種（Organism）での分類。②ヒトやマウスのデータが圧倒的に多いのが分かります



http://ddbj.nig.ac.jp//DRASearch/

DRA Search

DRASearch

Search Home | DRA Home

Accession :

Organism : StudyType :

CenterName : Platform :

Keyword :

Show 20 records Sort by Study Search Clear

Data Last Update 2018-04-17

Statistics

Released Entries

Type	Count
Submission	851945
Study	138718
Experiment	4006817
Sample	3602010
Run	[illegible]

①

Organism

#	Organism Name	Study
1	Homo sapiens	12984
2	Mus musculus	10491
3	soil metagenome	3875
4	marine metagenome	1680
5	Arabidopsis thaliana	1671
6	Panicum virgatum	1557
7	Drosophila melanogaster	1542
8	Oryza sativa	1502
9	Saccharomyces cerevisiae	1173
10	Populus trichocarpa	1146

②

Study Type

#	Study Type	Study
1	[illegible]le Genome Sequencing	49231
2	[illegible]er	49229
3	Metagenomics	19090
4	Transcriptome Analysis	19035
5	Population Genomics	791
6	Epigenetics	705
7	Exome Sequencing	248
8	Transcriptome Sequencing	170
9	Cancer Genomics	133
10	Pooled Clone Sequencing	35

Center Name

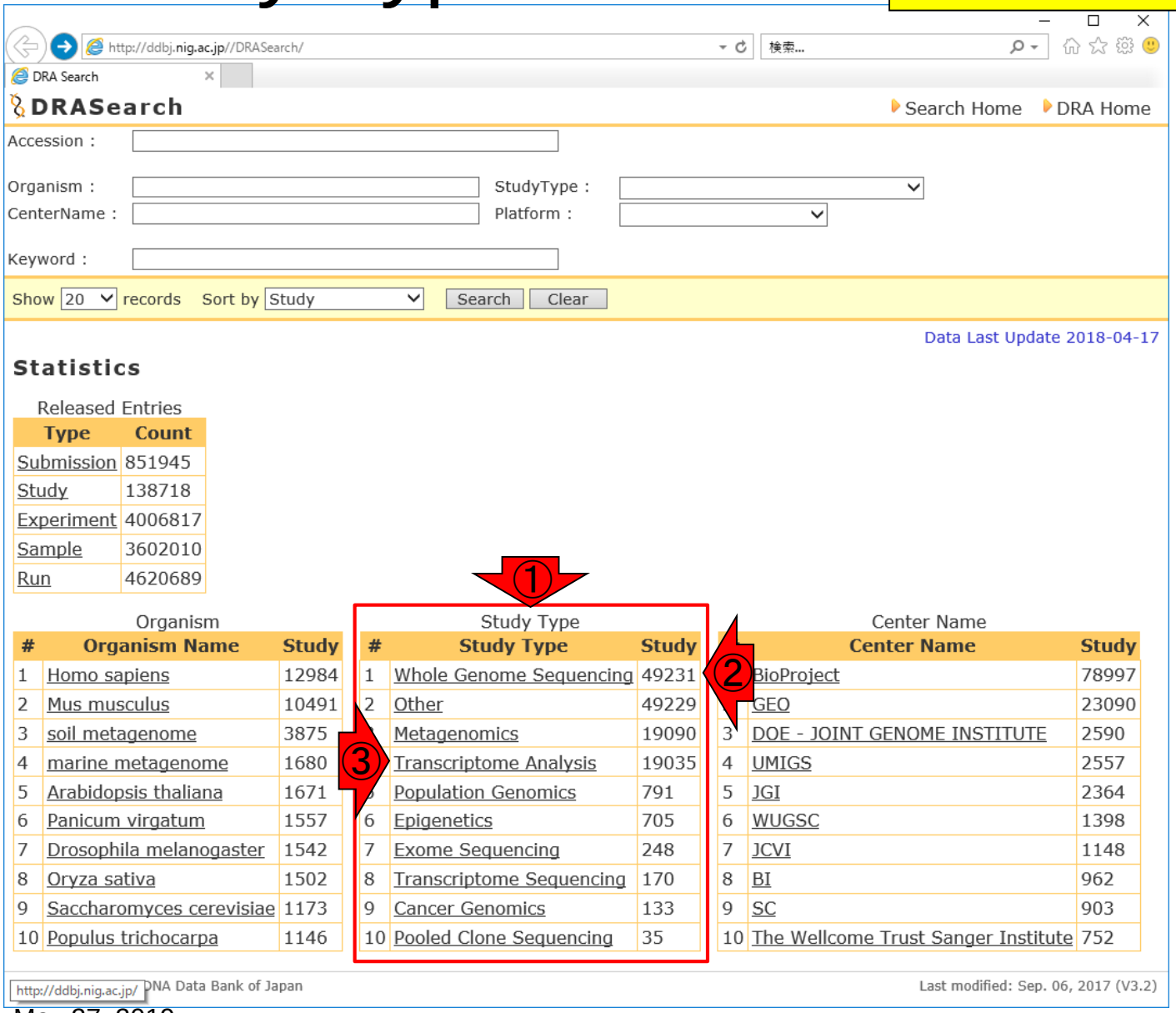
#	Center Name	Study
1	BioProject	78997
2	GEO	23090
3	DOE - JOINT GENOME INSTITUTE	2590
4	UMIGS	2557
5	JGI	2364
6	WUGSC	1398
7	JCVI	1148
8	BI	962
9	SC	903
10	The Wellcome Trust Sanger Institute	752

http://ddbj.nig.ac.jp/ [illegible]NA Data Bank of Japan

Last modified: Sep. 06, 2017 (V3.2)

Study Typeでの分類

①Study Typeでの分類。②最も多いのはゲノム配列決定、③トランスクリプトーム解析もそれなりにやられていることがわかります。③をクリック



http://ddbj.nig.ac.jp//DRASearch/

DRA Search

DRASearch

Search Home | DRA Home

Accession :

Organism : StudyType :

CenterName : Platform :

Keyword :

Show 20 records Sort by Study Search Clear

Data Last Update 2018-04-17

Statistics

Released Entries

Type	Count
Submission	851945
Study	138718
Experiment	4006817
Sample	3602010
Run	4620689

①

Organism

#	Organism Name	Study
1	Homo sapiens	12984
2	Mus musculus	10491
3	soil metagenome	3875
4	marine metagenome	1680
5	Arabidopsis thaliana	1671
6	Panicum virgatum	1557
7	Drosophila melanogaster	1542
8	Oryza sativa	1502
9	Saccharomyces cerevisiae	1173
10	Populus trichocarpa	1146

Study Type

#	Study Type	Study
1	Whole Genome Sequencing	49231
2	Other	49229
3	Metagenomics	19090
4	Transcriptome Analysis	19035
5	Population Genomics	791
6	Epigenetics	705
7	Exome Sequencing	248
8	Transcriptome Sequencing	170
9	Cancer Genomics	133
10	Pooled Clone Sequencing	35

②

Center Name

#	Center Name	Study
1	BioProject	78997
2	GEO	23090
3	DOE - JOINT GENOME INSTITUTE	2590
4	UMIGS	2557
5	JGI	2364
6	WUGSC	1398
7	JCVI	1148
8	BI	962
9	SC	903
10	The Wellcome Trust Sanger Institute	752

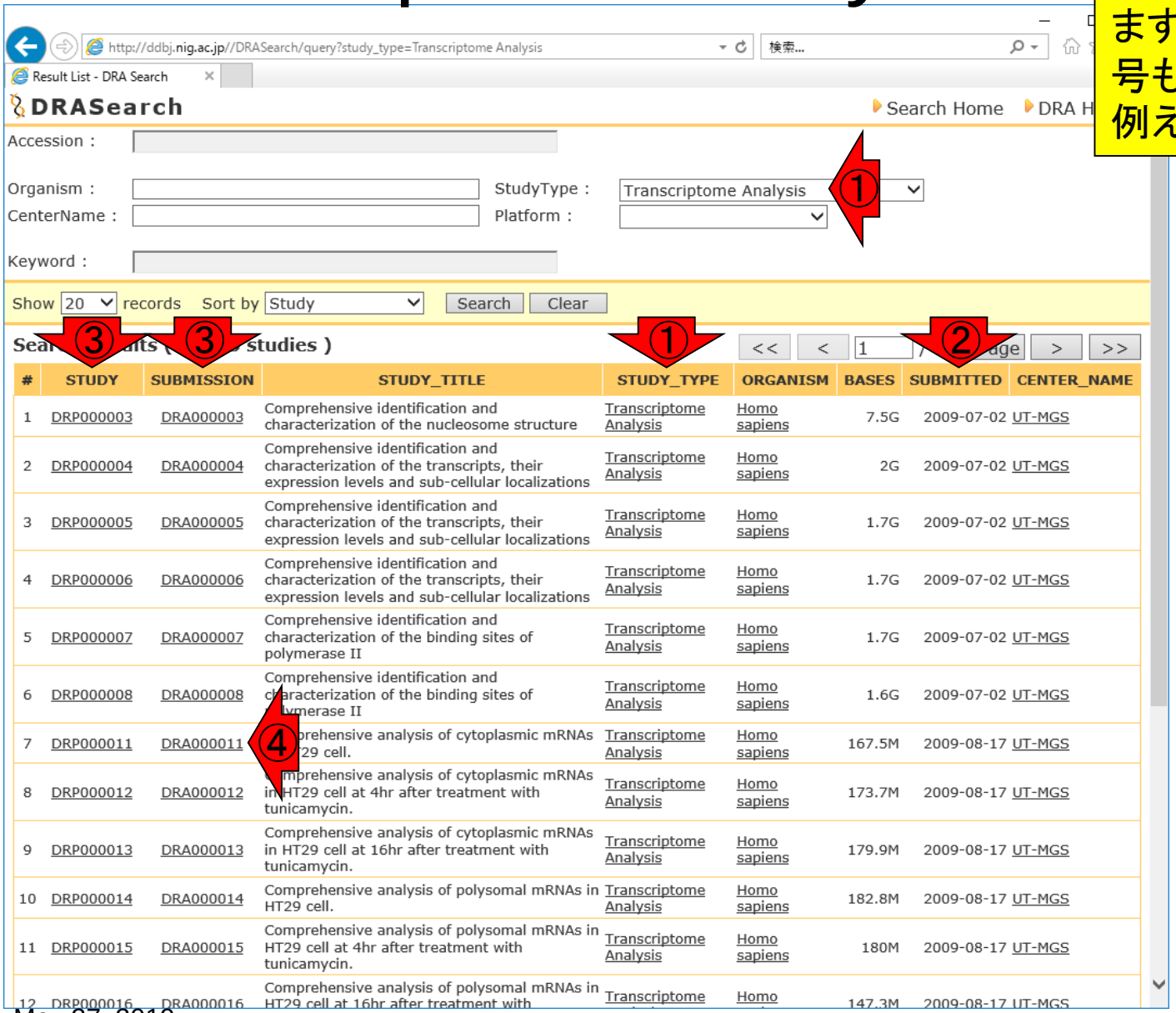
③

http://ddbj.nig.ac.jp/ DNA Data Bank of Japan

Last modified: Sep. 06, 2017 (V3.2)

Transcriptome Analysis

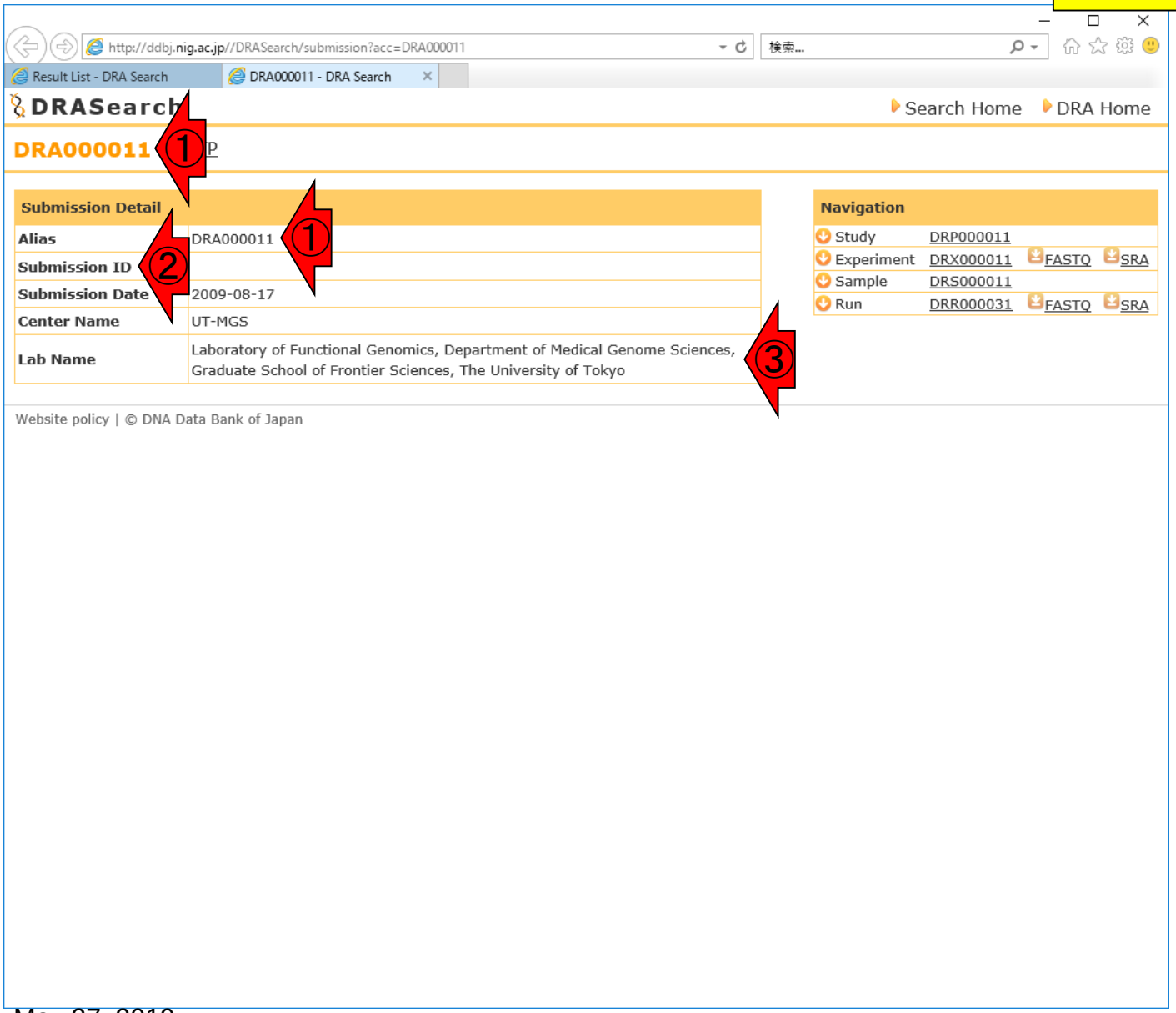
①がTranscriptome Analysisになりました。②DDBJへの登録日順（古→新）になっていることがわかります。必然的に、③IDのシリアル番号も一桁台の数字が見られます。例えば④のDRA000011をクリック



#	STUDY	SUBMISSION	STUDY_TITLE	STUDY_TYPE	ORGANISM	BASES	SUBMITTED	CENTER_NAME
1	DRP000003	DRA000003	Comprehensive identification and characterization of the nucleosome structure	Transcriptome Analysis	Homo sapiens	7.5G	2009-07-02	UT-MGS
2	DRP000004	DRA000004	Comprehensive identification and characterization of the transcripts, their expression levels and sub-cellular localizations	Transcriptome Analysis	Homo sapiens	2G	2009-07-02	UT-MGS
3	DRP000005	DRA000005	Comprehensive identification and characterization of the transcripts, their expression levels and sub-cellular localizations	Transcriptome Analysis	Homo sapiens	1.7G	2009-07-02	UT-MGS
4	DRP000006	DRA000006	Comprehensive identification and characterization of the transcripts, their expression levels and sub-cellular localizations	Transcriptome Analysis	Homo sapiens	1.7G	2009-07-02	UT-MGS
5	DRP000007	DRA000007	Comprehensive identification and characterization of the binding sites of polymerase II	Transcriptome Analysis	Homo sapiens	1.7G	2009-07-02	UT-MGS
6	DRP000008	DRA000008	Comprehensive identification and characterization of the binding sites of polymerase II	Transcriptome Analysis	Homo sapiens	1.6G	2009-07-02	UT-MGS
7	DRP000011	DRA000011	Comprehensive analysis of cytoplasmic mRNAs in HT29 cell.	Transcriptome Analysis	Homo sapiens	167.5M	2009-08-17	UT-MGS
8	DRP000012	DRA000012	Comprehensive analysis of cytoplasmic mRNAs in HT29 cell at 4hr after treatment with tunicamycin.	Transcriptome Analysis	Homo sapiens	173.7M	2009-08-17	UT-MGS
9	DRP000013	DRA000013	Comprehensive analysis of cytoplasmic mRNAs in HT29 cell at 16hr after treatment with tunicamycin.	Transcriptome Analysis	Homo sapiens	179.9M	2009-08-17	UT-MGS
10	DRP000014	DRA000014	Comprehensive analysis of polysomal mRNAs in HT29 cell.	Transcriptome Analysis	Homo sapiens	182.8M	2009-08-17	UT-MGS
11	DRP000015	DRA000015	Comprehensive analysis of polysomal mRNAs in HT29 cell at 4hr after treatment with tunicamycin.	Transcriptome Analysis	Homo sapiens	180M	2009-08-17	UT-MGS
12	DRP000016	DRA000016	Comprehensive analysis of polysomal mRNAs in HT29 cell at 16hr after treatment with	Transcriptome	Homo	147.3M	2009-08-17	UT-MGS

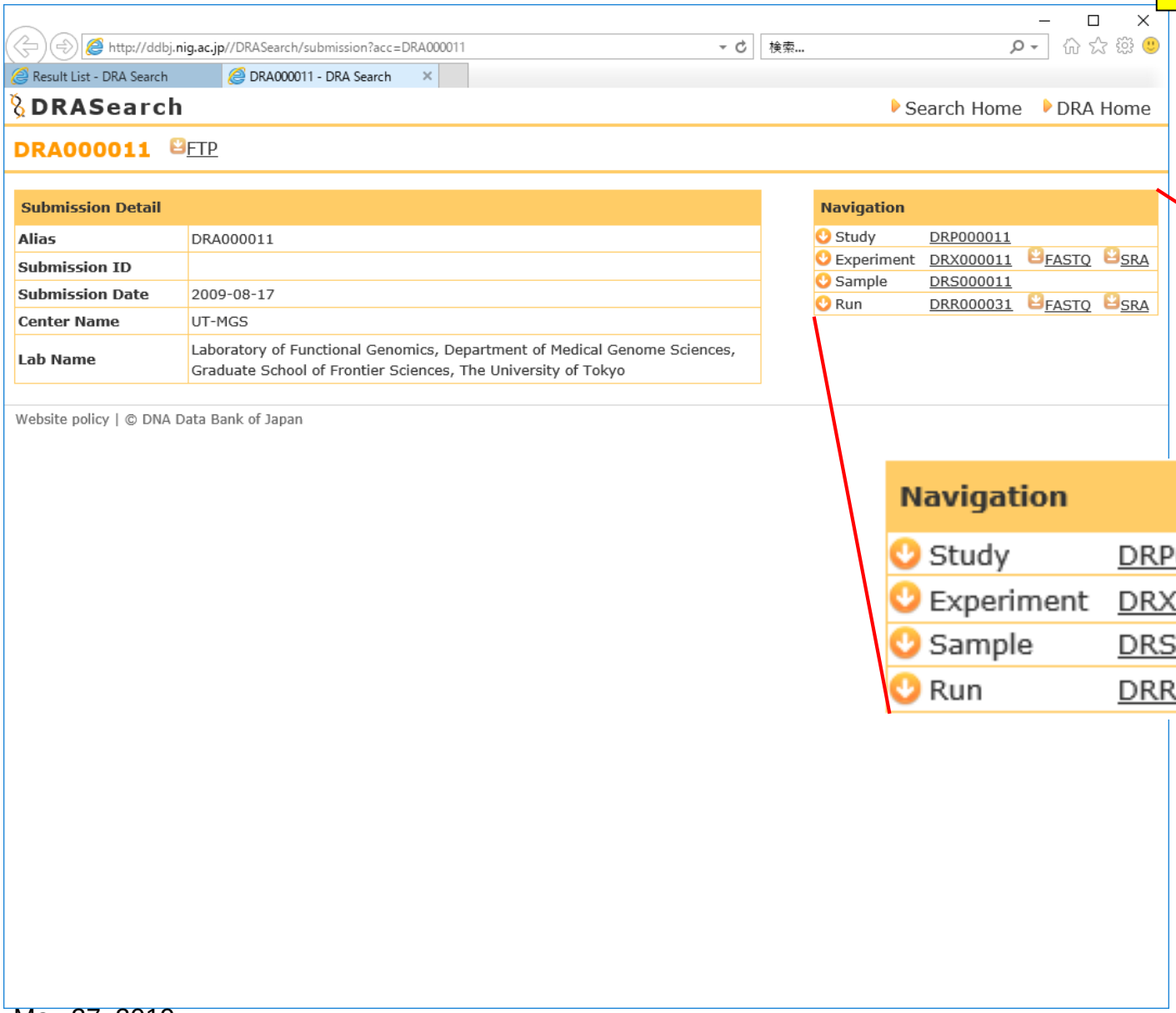
DRA000011

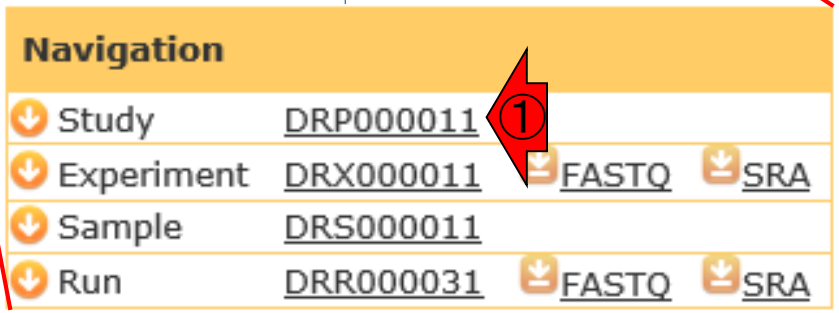
①DRA…は、②Submission ID。データ登録時に付与されるもので③、データ登録者に関する情報が含まれる



DRP000011

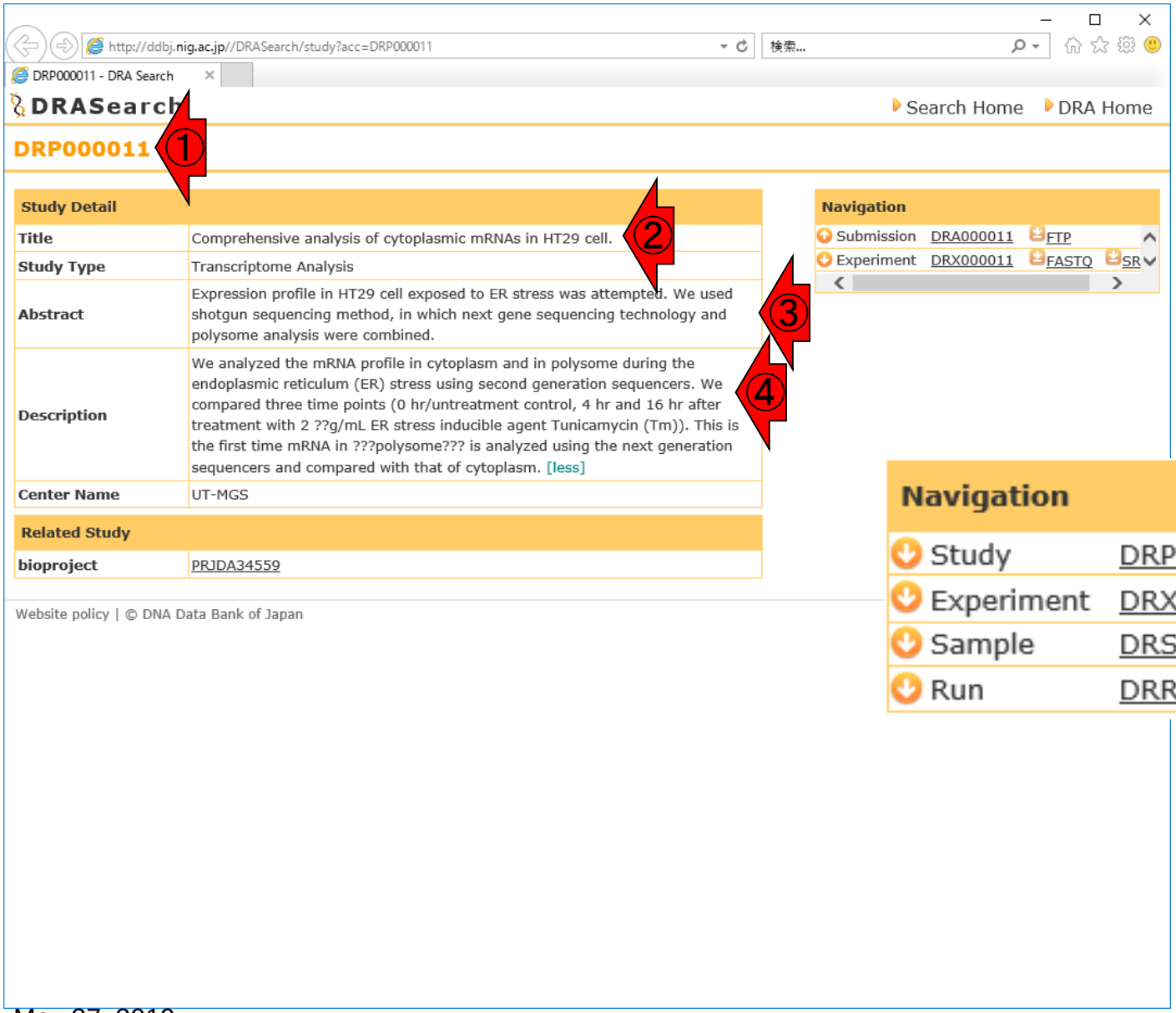
①DRP…はStudy ID。この研究に関する情報であり…①のリンク先を次のスライドで示す

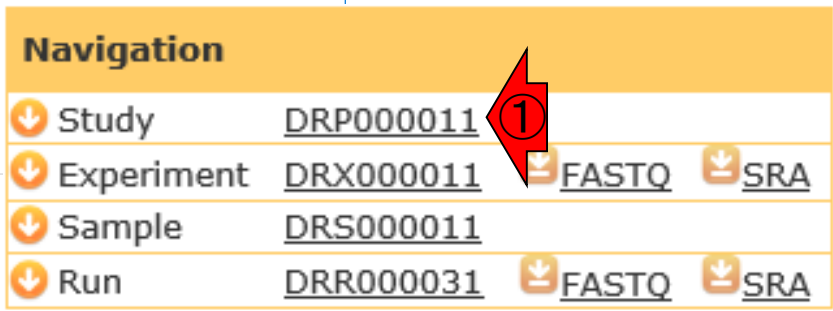




DRP000011

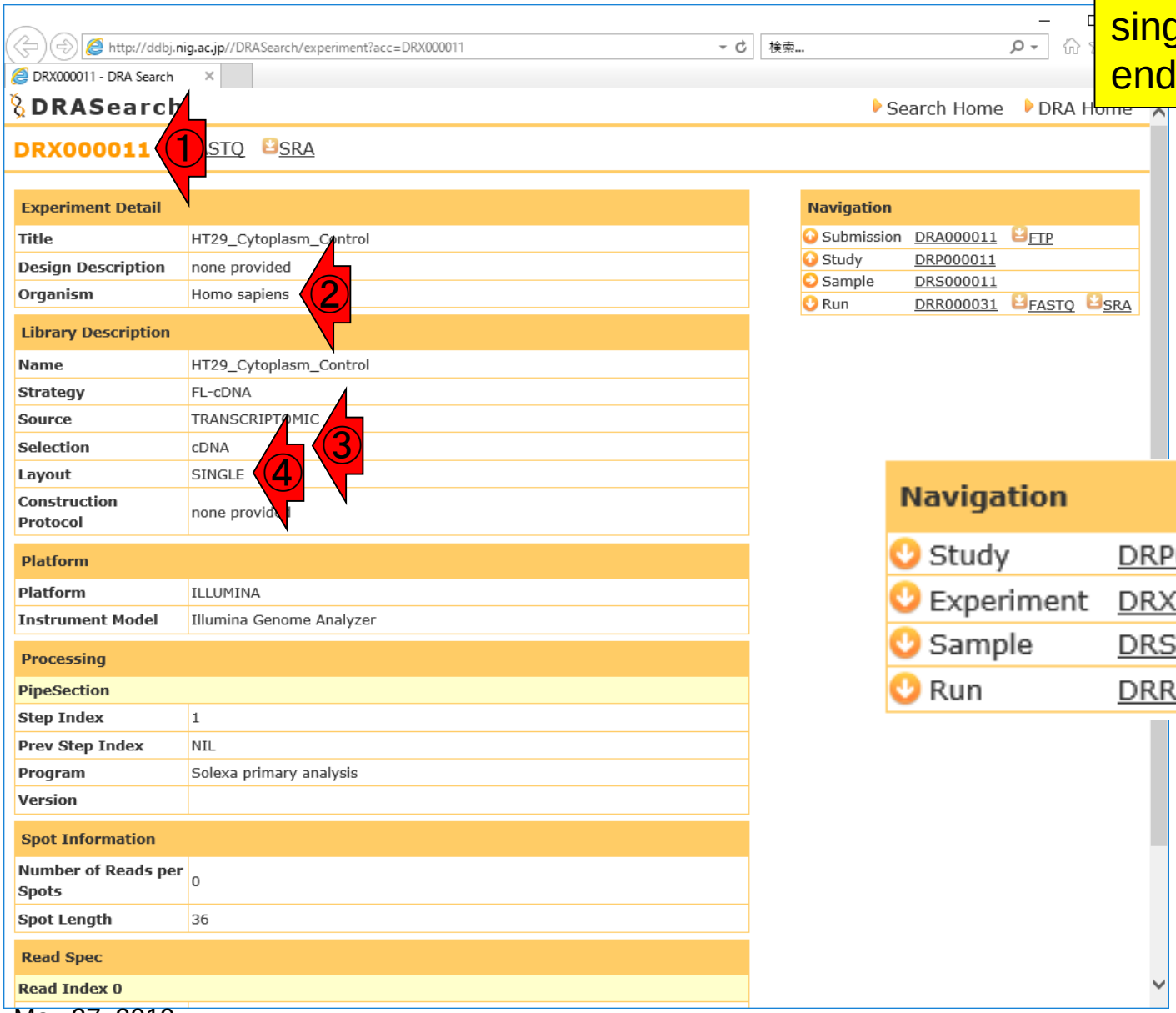
②タイトル、③Abstract、④descriptionなどの情報からなる

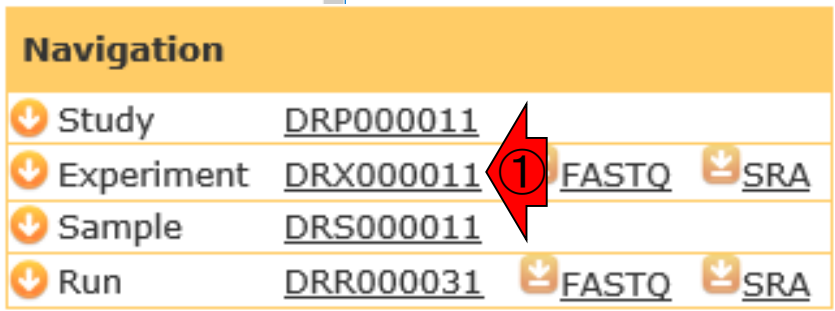




DRX000011

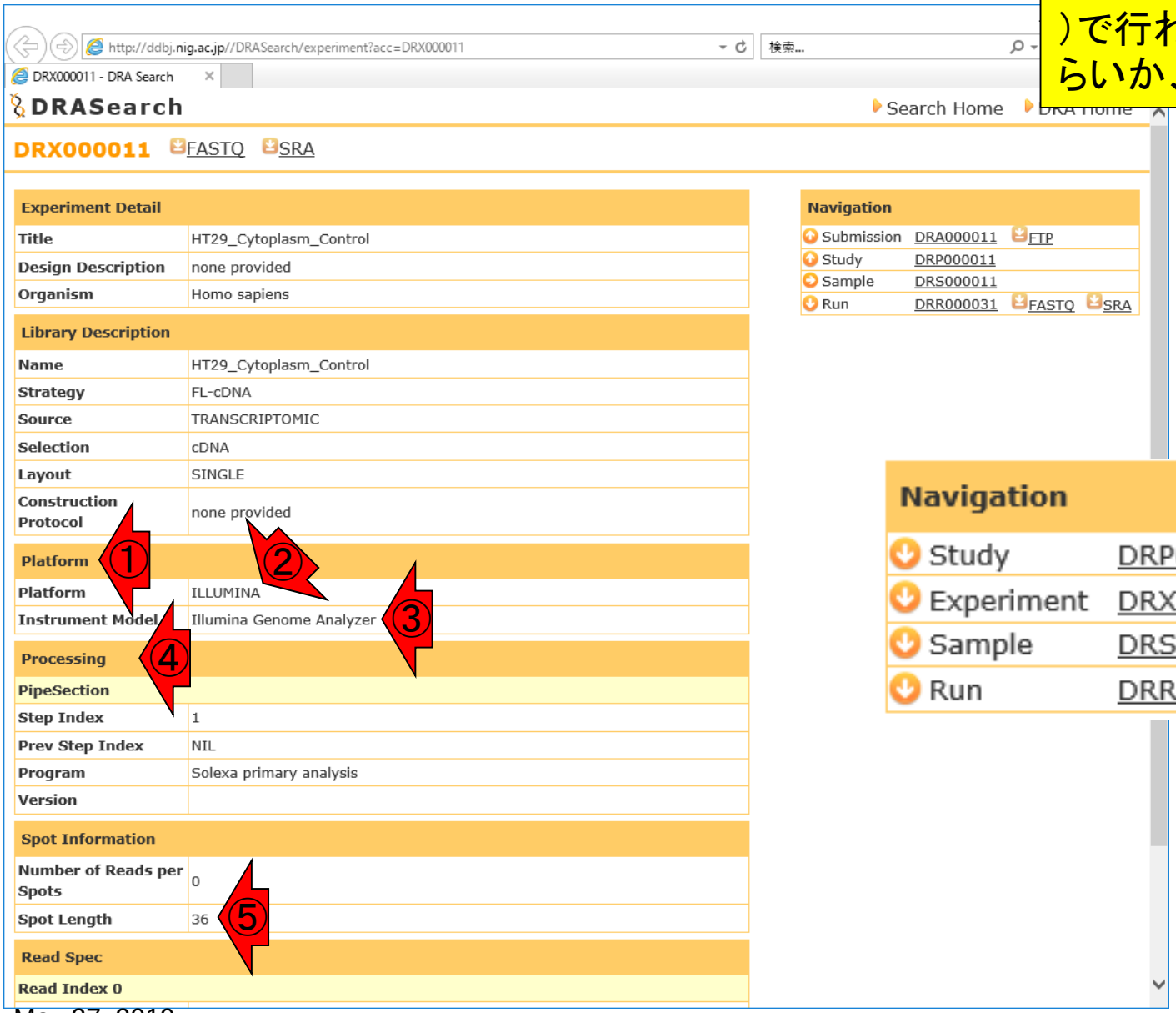
①DRX…はExperiment ID。実験情報が記載されており、②ヒトのデータであり、③cDNAであり、④single-endデータである（paired-endではない）ことがわかります





DRX000011

①プラットフォーム情報（②どのメーカーの、③どのNGS機器で取得されたか）、④どのような処理手順（プロトコル）で行われたのか、⑤リード長はどれくらいか、などの情報が含まれる



DRX000011

ちなみに、①Genome Analyzerという機種はNGS機器の中でも相当古いものであり、データ登録時の2009年頃はまだ使われていた。NGSが出始めの頃は、②36塩基程度しか読めなかった。それが、**ショートリード**と評されていた所以です

http://ddbj.nig.ac.jp//DRASearch/experiment?acc=DRX000011

DRX000011 - DRA Search

DRASearch

DRX000011 FASTQ SRA

Experiment Detail	
Title	HT29_Cytoplasm_Control
Design Description	none provided
Organism	Homo sapiens

Library Description	
Name	HT29_Cytoplasm_Control
Strategy	FL-cDNA
Source	TRANSCRIPTOMIC
Selection	cDNA
Layout	SINGLE
Construction Protocol	none provided

Platform	
Platform	ILLUMINA
Instrument Model	Illumina Genome Analyzer

Processing	
PipeSection	
Step Index	1
Prev Step Index	NIL
Program	Solexa primary analysis
Version	

Spot Information	
Number of Reads per Spots	0
Spot Length	36

Read Spec
Read Index 0

①

②

Navigation			
Submission	DRA000011	FTP	
Study	DRP000011		
Sample	DRS000011		
Run	DRR000031	FASTQ	SRA

Navigation

Study	DRP000011		
Experiment	DRX000011	FASTQ	SRA
Sample	DRS000011		
Run	DRR000031	FASTQ	SRA

DRS000011

①DRS…はSample ID。このデータの場合は、情報量が少ないですね。②のリンク先にいけばわかりますが、ヒトのサンプルだということしかわかりません

http://ddbj.nig.ac.jp//DRASearch/sample?acc=DRS000011

DRS000011 - DRA Search

DRASearch

Search Home | DRA Home

DRS000011

Sample Detail	
Title	DRS000011
Description	none provided

Organism Info	
Taxon ID	9606
Common Name	
Scientific Name	Homo sapiens
Anonymized Name	
Individual Name	

②

Navigation			
Submission	DRA000011	FTP	
Study	DRP000011		
Experiment	DRX000011	FASTQ	SRA

Website policy | © DNA Data Bank of Japan

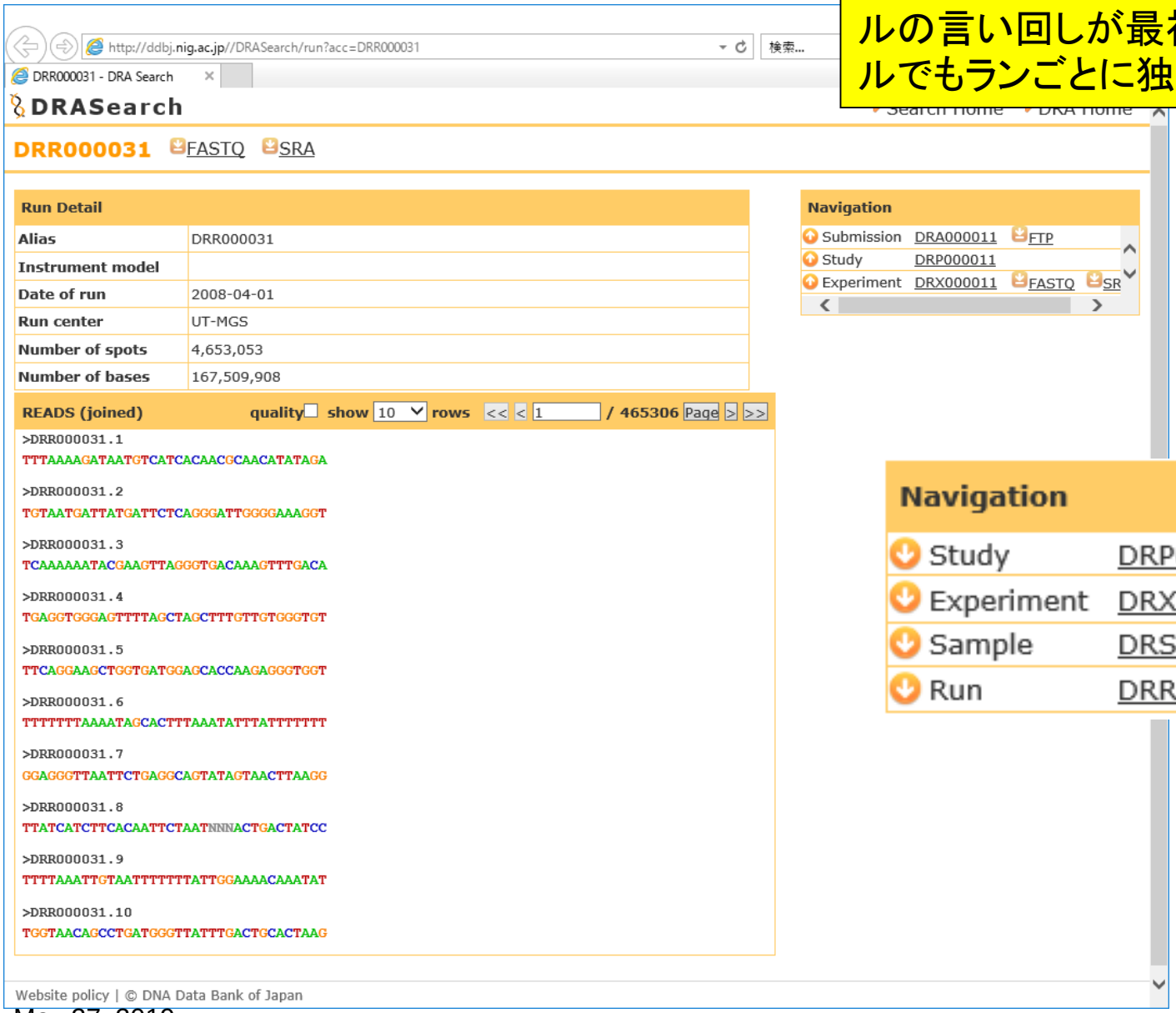
Navigation

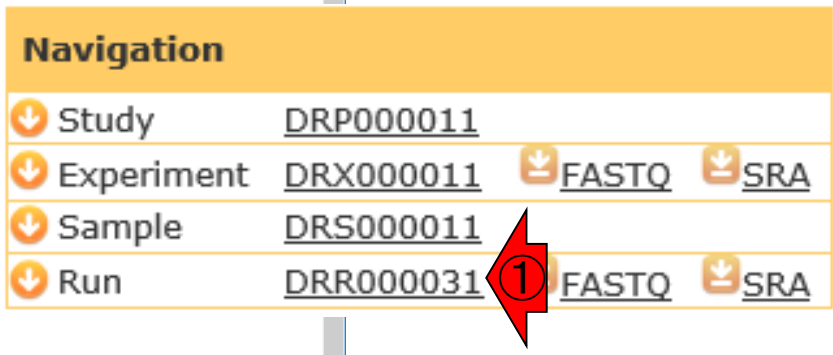
Study	DRP000011		
Experiment	DRX000011	FASTQ	SRA
Sample	DRS000011		
Run	DRR000031	FASTQ	SRA

①

DRR000031

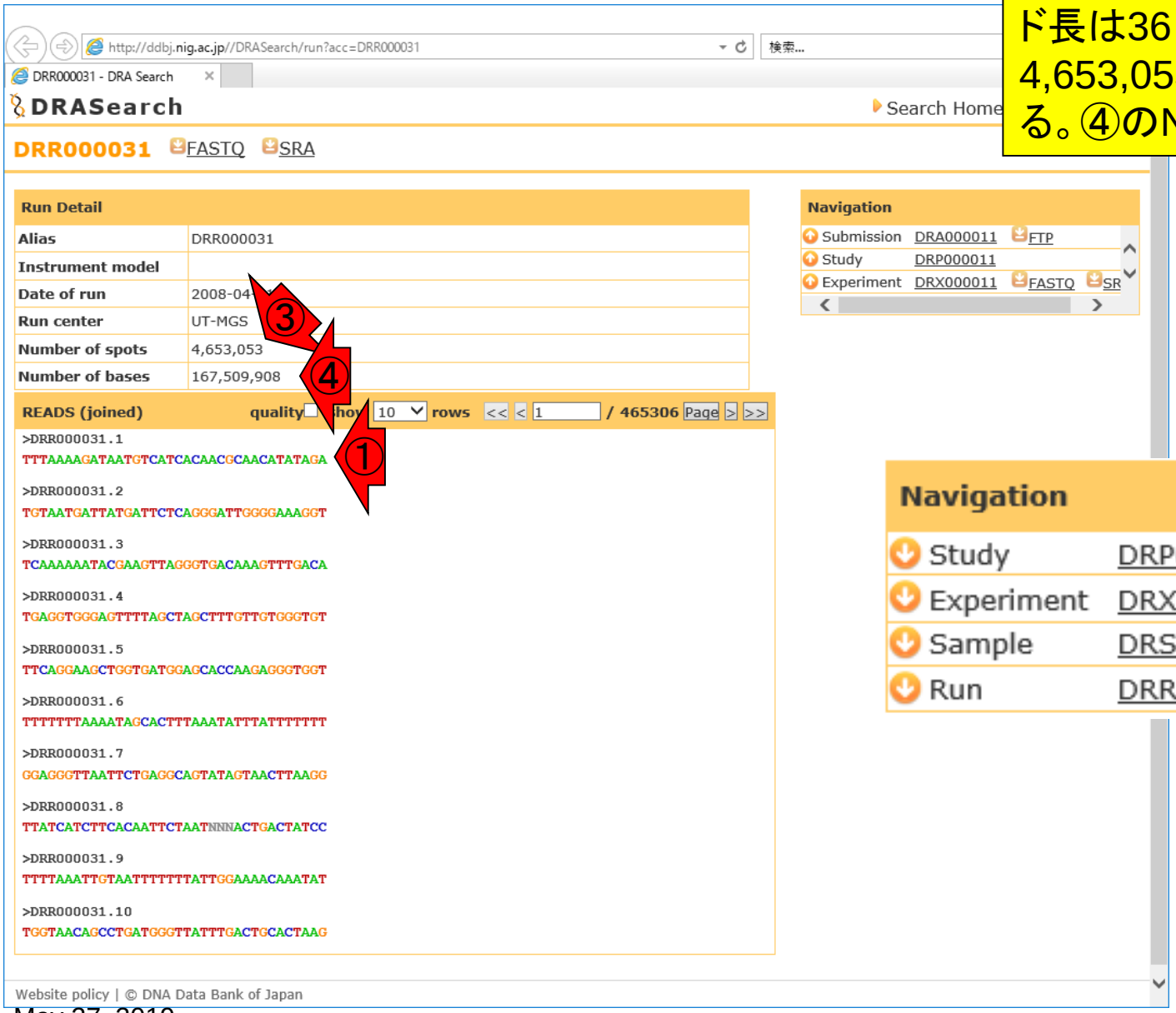
NGSの場合は、実験の単位をラン（Run）といいます。これはNGS分野で大きなシェアを占めるIllumina（によって買収されたSolexa社）のプロトコルの言い回しが最初だったと思います。同じサンプルでもランごとに独立のIDが付与されます

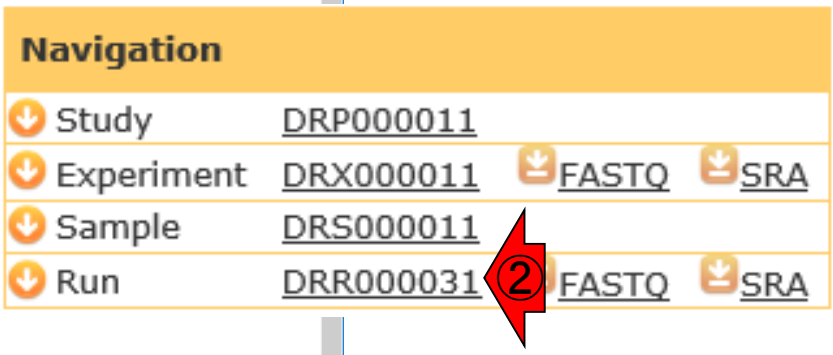




DRR000031

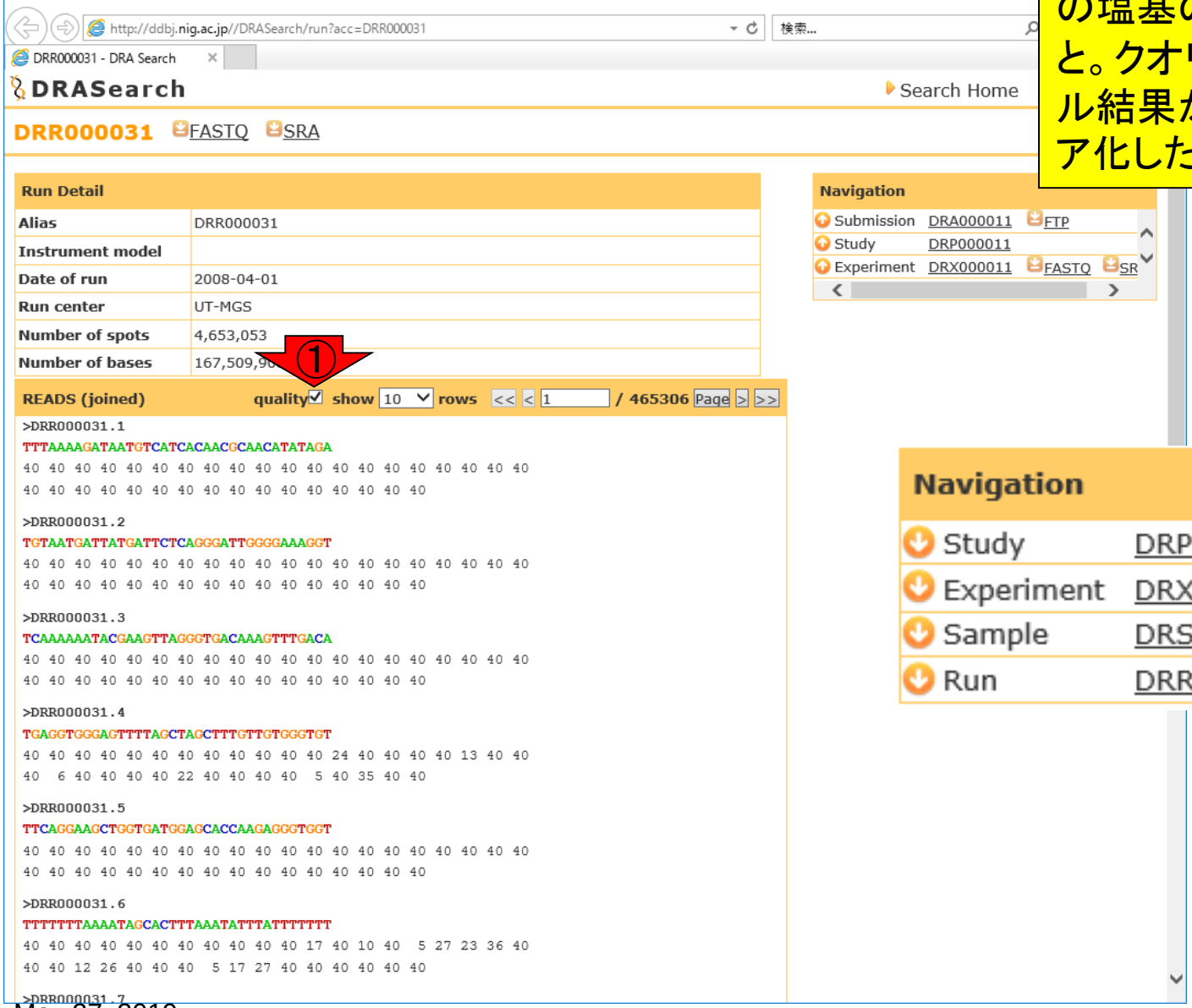
①これがリードの実体。②DRR000031の場合は、総リード数が③4,653,053個あることがわかる（約465万リード）。リード長は36 bpなので、総塩基数が4,653,053×36 = 167,509,908 bpとなる。④のNumber of basesと完全に一致





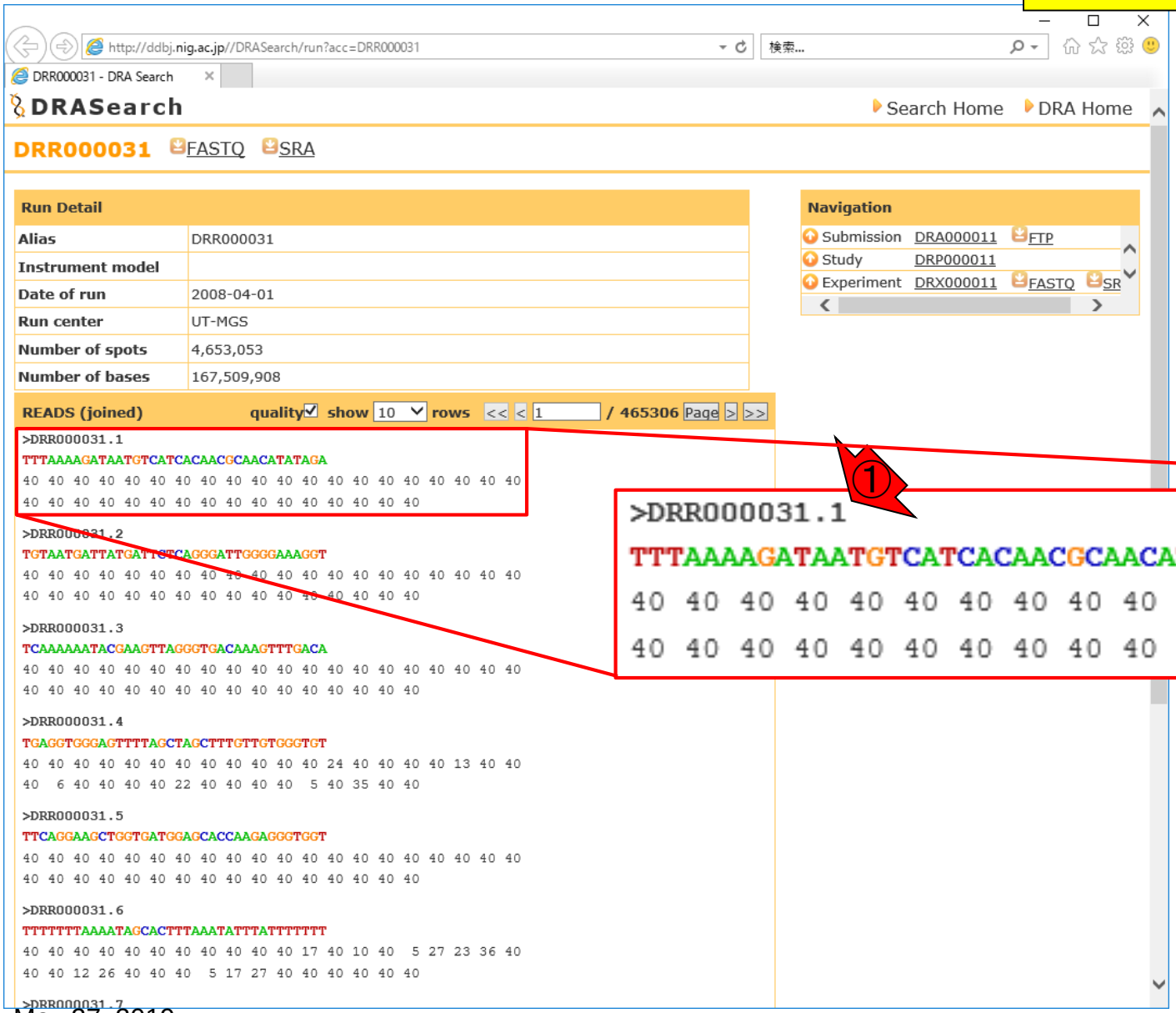
クオリティ情報

①qualityのところにチェックをいれると、クオリティスコア情報も表示される。ベースコールとは、A,C,G,Tからなる4文字の塩基のうち、どれか1つを選択すること。クオリティスコアは、そのベースコール結果がどれだけ確からしいかをスコア化したものであり、高いほどよい



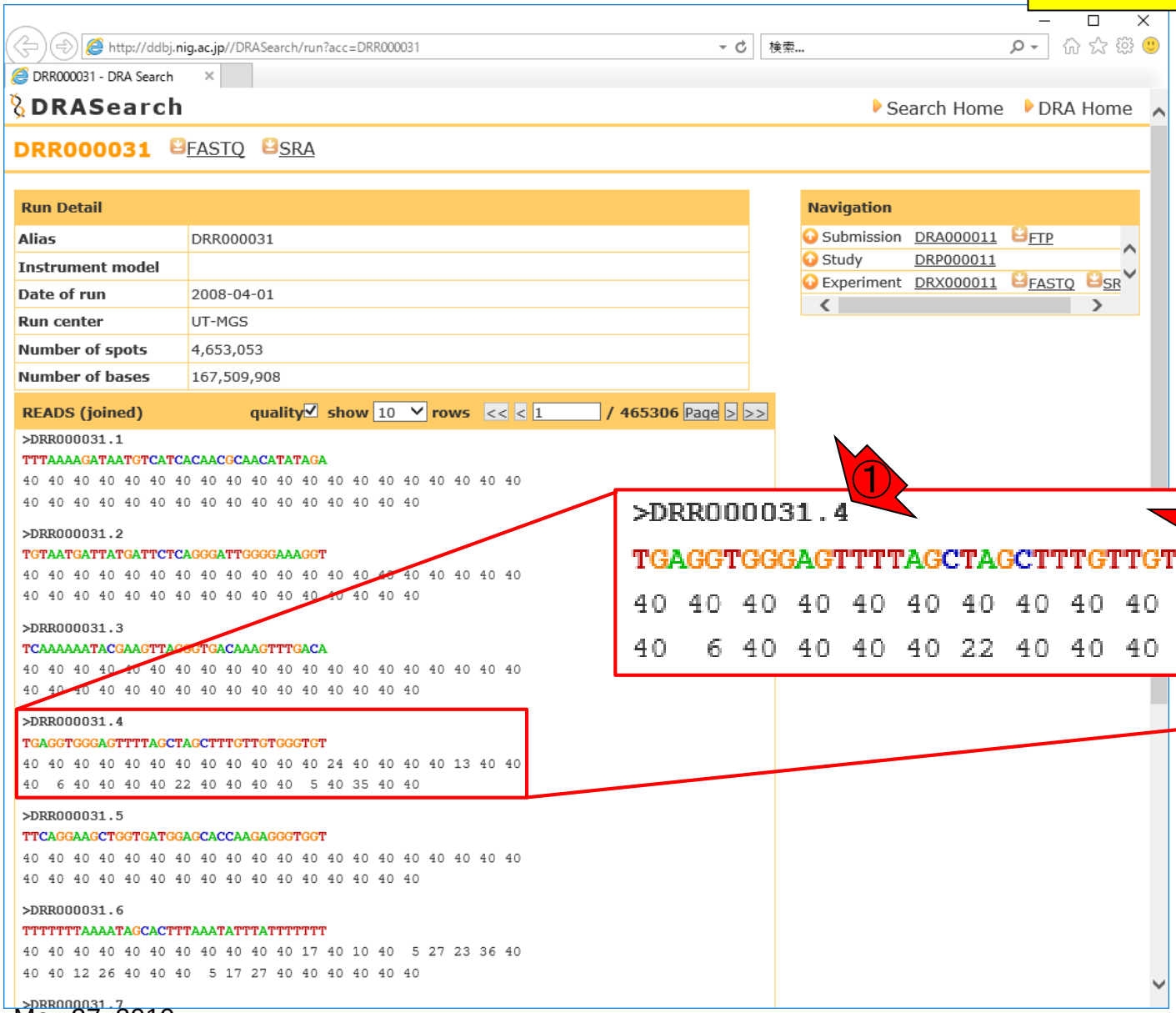
クオリティ情報

例えば、①一番最初のリード（リードIDがDRR000031.1）の②最後の塩基のクオリティスコアは、③40と読み解きます



クオリティ情報

また、①4番目のリード（DRR000031.4）の、②右から5番目の塩基のクオリティスコアは、③5と読み解きます



ベースコールエラー率

クオリティスコアqの閾値は、20や30が目安。q = 20はベースコール結果が間違っている確率（エラー率p）が1%という意味である。また、q = 30はp = 0.1%に相当する

http://ddbj.nig.ac.jp//DRASearch/run?acc=DRR000031

DRR000031 - DRA Search

DRASearch　Search Home　DRA Home

DRR000031 FASTQ SRA

Run Detail	
Alias	DRR000031
Instrument model	
Date of run	2008-04-01
Run center	UT-MGS
Number of spots	4,653,053
Number of bases	167,509,908

Navigation

Submission	DRA000011	FTP
Study	DRP000011	
Experiment	DRX000011	FASTQ SR

READS (joined)　quality ☑ show 10 rows　<< < 1 / 465306 Page > >>

```
>DRR000031.1
TTTAAAAGATAATGTCATCACAACGCAACATATAGA
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40

>DRR000031.2
TGTAATGATTATGATTCTCAGGGATTGGGGAAAGGT
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40

>DRR000031.3
TCAAAAAATACGAAGTTAGGGTGACAAAGTTTGACA
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40

>DRR000031.4
TGAGGTGGGAGTTTTAGCTAGCTTTGTTGTGGGTGT
40 40 40 40 40 40 40 40 40 40 40 24 40 40 40 40 13 40 40
40  6 40 40 40 22 40 40 40 40  5 40 35 40 40

>DRR000031.5
TTCAGGAAGCTGGTGATGGAGCACCAAGAGGGTGGT
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40

>DRR000031.6
TTTTTTTAAAATAGCACTTTAAATATTTATTTTTTT
40 40 40 40 40 40 40 40 40 40 40 17 40 10 40  5 27 23 36 40
40 40 12 26 40 40 40  5 17 27 40 40 40 40 40 40

>DRR000031.7
```


>DRR000031.4

TGAGGTGGGAGTTTTAGCTAGCTTTGTTGTGGGTGT

40 40 40 40 40 40 40 40 40 40 40 40 24 40 40 40 40 13 40 40

40 6 40 40 40 40 22 40 40 40 40 5 40 35 40 40

数式で表すと…

$$q = -10 \times log_{10}(p)$$ ①

$$q = -10 \times log_{10}(10^{-3})$$ ②
$$q = -10 \times (-3) = 30$$

$$q = -10 \times log_{10}(10^{-5})$$ ③
$$q = -10 \times (-5) = 50$$

クオリティスコアqとエラー率pの関係は、①式で表されます。一見ややこしいですが、②p = 0.1% = 10^{-3}だと考えれば意外と簡単です。③エラー率が低いほどクオリティスコアqは上がります。

数式で表すと…

$$q = -10 \times log_{10}(p)$$ ①

$$q = -10 \times log_{10}(10^{-3})$$ ②
$$q = -10 \times (-3) = 30$$

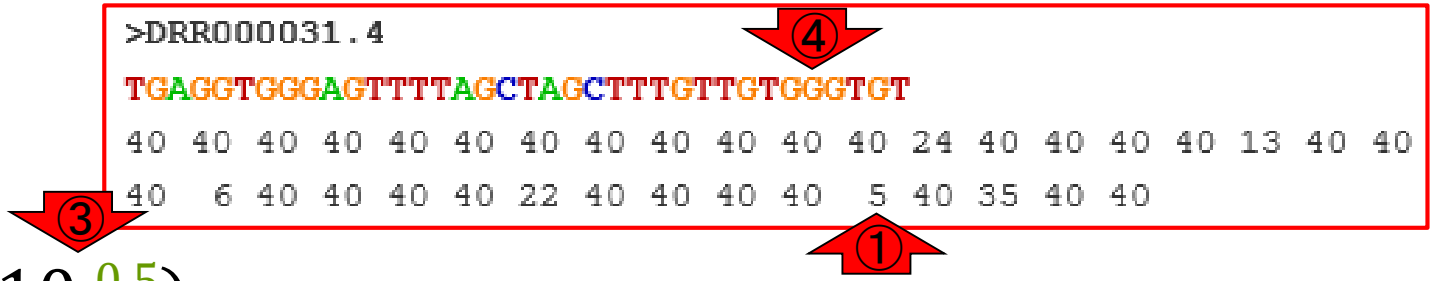
$$q = -10 \times log_{10}(10^{-5})$$ ③
$$q = -10 \times (-5) = 50$$

$$q = -10 \times log_{10}(10^{-1})$$ ④
$$q = -10 \times (-1) = 10$$

クオリティスコアqとエラー率pの関係は、①式で表されます。一見ややこしいですが、②$p = 0.1\% = 10^{-3}$だと考えれば意外と簡単です。③エラー率が低いほどクオリティスコアqは上がります。④エラー率が高いほどクオリティスコアqは下がります。

クオリティスコアq = 5の場合

①クオリティスコアqが5の場合は、②が-0.5になるので、③エラー率p = $10^{-0.5}$= 0.316となる。④Gというベースコール結果は正確性が低いと判断する



$$q = -10 \times log_{10}(10^{-0.5})$$

$$q = -10 \times (-0.5) = 5$$

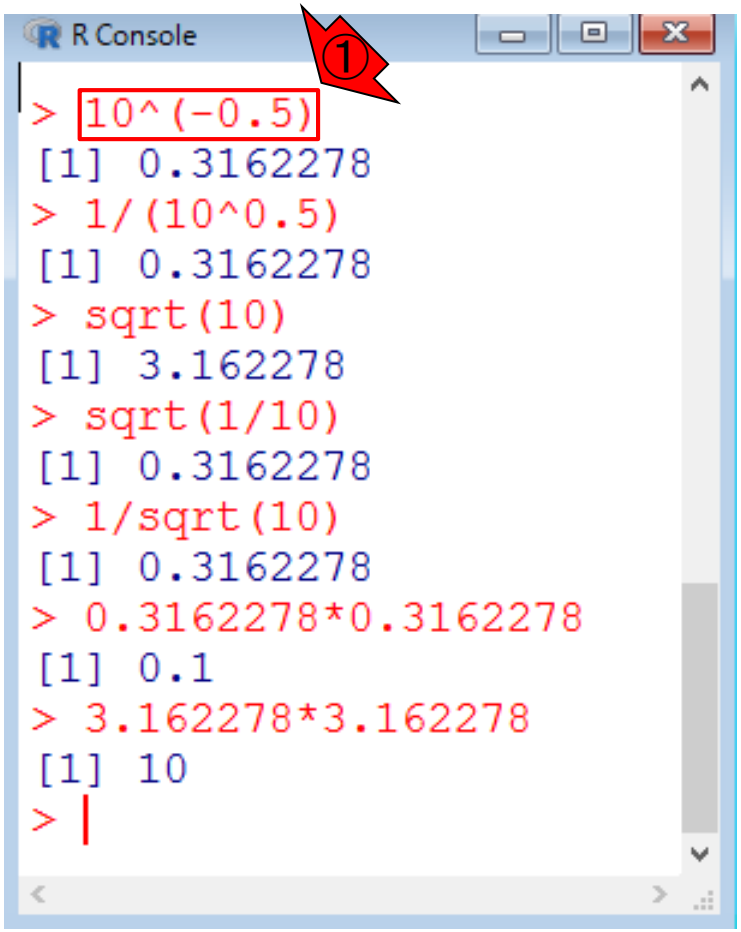
②　①

Tips

①こういうことです。2乗して10になるのが3.162278くらいであることを思い出せれば、なんとか理解できるでしょう

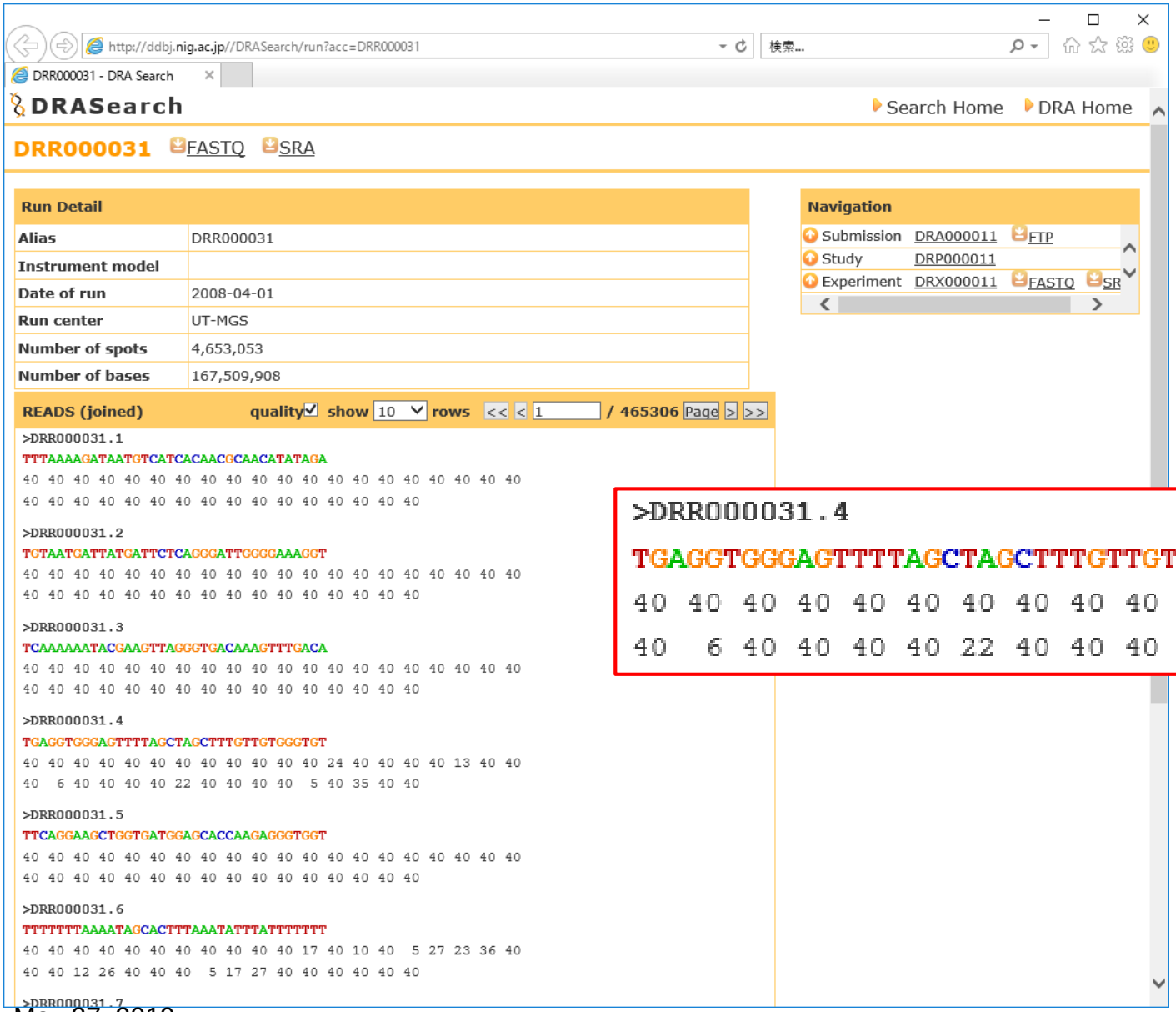
$$q = -10 \times log_{10}(10^{-0.5})$$
$$q = -10 \times (-0.5) = 5$$

①



おさらい

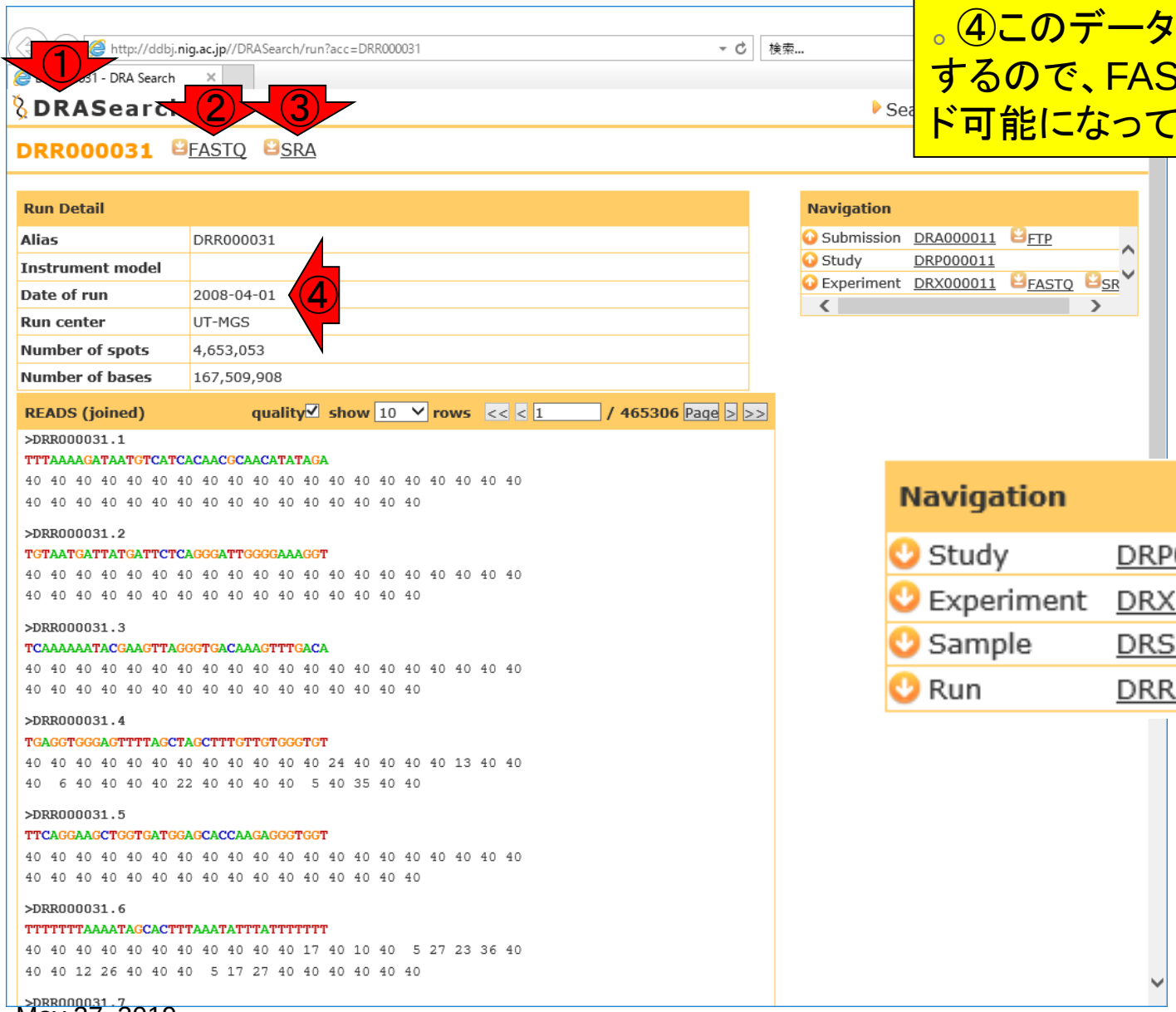
クオリティスコアqの閾値は、
キリがいいので20や30が目安

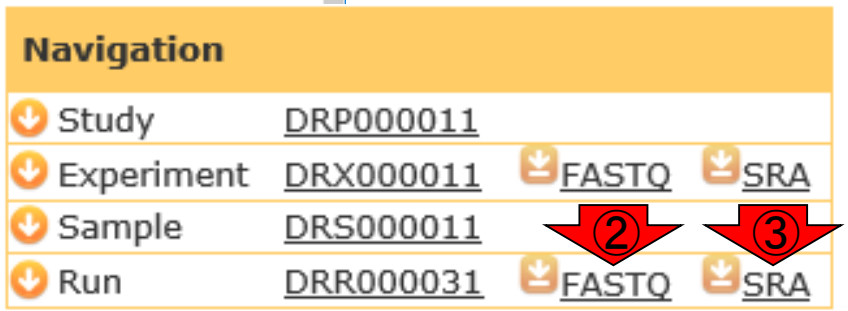


>DRR000031.4
TGAGGTGGGAGTTTTAGCTAGCTTTGTTGTGGGTGT
40 40 40 40 40 40 40 40 40 40 40 40 24 40 40 40 40 13 40 40
40 6 40 40 40 40 22 40 40 40 40 5 40 35 40 40

データのダウンロード

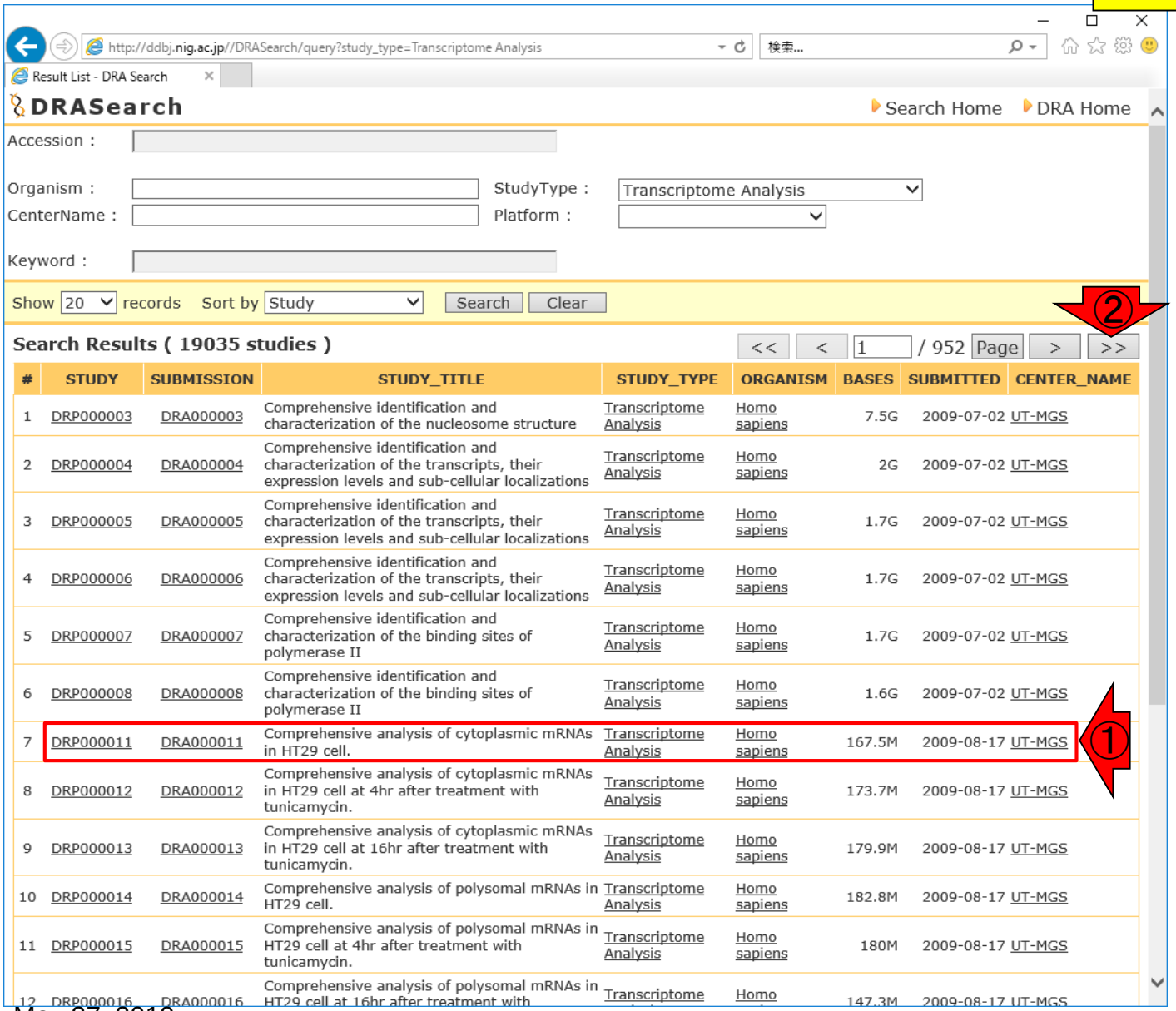
①DRAの場合は、②FASTQ形式、③SRA形式ファイルのいずれでもダウンロード可能。同じ番号のところなら、どちらをクリックしてもよい。④このデータは10年以上前のものから存在するので、FASTQとSRAの両方がダウンロード可能になっている





最新のデータだと…

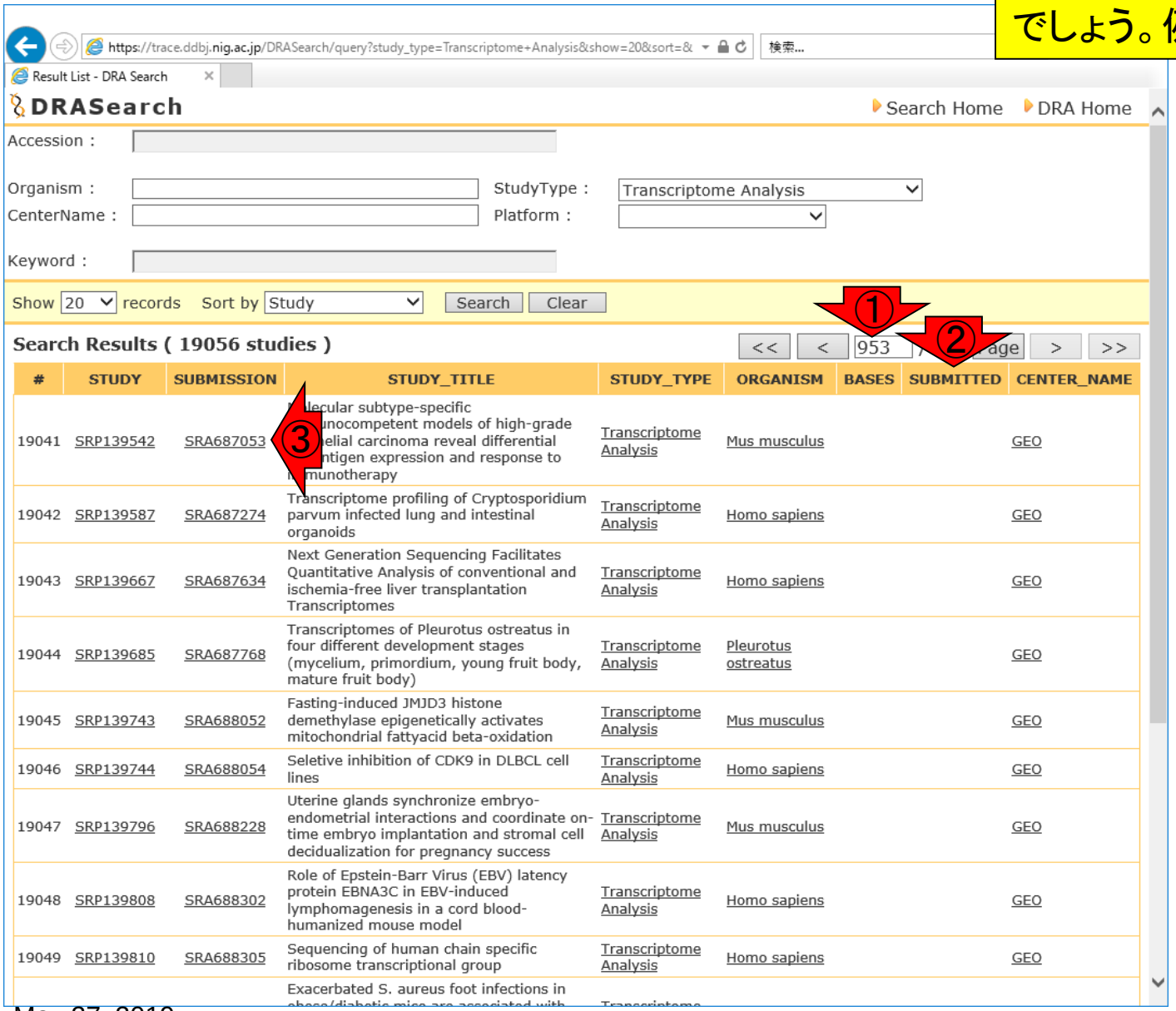
これまで見ていたのは、①のデータなので、②を押して最新のデータがあると思われる最終ページに飛ぶ



#	STUDY	SUBMISSION	STUDY_TITLE	STUDY_TYPE	ORGANISM	BASES	SUBMITTED	CENTER_NAME
1	DRP000003	DRA000003	Comprehensive identification and characterization of the nucleosome structure	Transcriptome Analysis	Homo sapiens	7.5G	2009-07-02	UT-MGS
2	DRP000004	DRA000004	Comprehensive identification and characterization of the transcripts, their expression levels and sub-cellular localizations	Transcriptome Analysis	Homo sapiens	2G	2009-07-02	UT-MGS
3	DRP000005	DRA000005	Comprehensive identification and characterization of the transcripts, their expression levels and sub-cellular localizations	Transcriptome Analysis	Homo sapiens	1.7G	2009-07-02	UT-MGS
4	DRP000006	DRA000006	Comprehensive identification and characterization of the transcripts, their expression levels and sub-cellular localizations	Transcriptome Analysis	Homo sapiens	1.7G	2009-07-02	UT-MGS
5	DRP000007	DRA000007	Comprehensive identification and characterization of the binding sites of polymerase II	Transcriptome Analysis	Homo sapiens	1.7G	2009-07-02	UT-MGS
6	DRP000008	DRA000008	Comprehensive identification and characterization of the binding sites of polymerase II	Transcriptome Analysis	Homo sapiens	1.6G	2009-07-02	UT-MGS
7	DRP000011	DRA000011	Comprehensive analysis of cytoplasmic mRNAs in HT29 cell.	Transcriptome Analysis	Homo sapiens	167.5M	2009-08-17	UT-MGS
8	DRP000012	DRA000012	Comprehensive analysis of cytoplasmic mRNAs in HT29 cell at 4hr after treatment with tunicamycin.	Transcriptome Analysis	Homo sapiens	173.7M	2009-08-17	UT-MGS
9	DRP000013	DRA000013	Comprehensive analysis of cytoplasmic mRNAs in HT29 cell at 16hr after treatment with tunicamycin.	Transcriptome Analysis	Homo sapiens	179.9M	2009-08-17	UT-MGS
10	DRP000014	DRA000014	Comprehensive analysis of polysomal mRNAs in HT29 cell.	Transcriptome Analysis	Homo sapiens	182.8M	2009-08-17	UT-MGS
11	DRP000015	DRA000015	Comprehensive analysis of polysomal mRNAs in HT29 cell at 4hr after treatment with tunicamycin.	Transcriptome Analysis	Homo sapiens	180M	2009-08-17	UT-MGS
12	DRP000016	DRA000016	Comprehensive analysis of polysomal mRNAs in HT29 cell at 16hr after treatment with	Transcriptome	Homo	147.3M	2009-08-17	UT-MGS

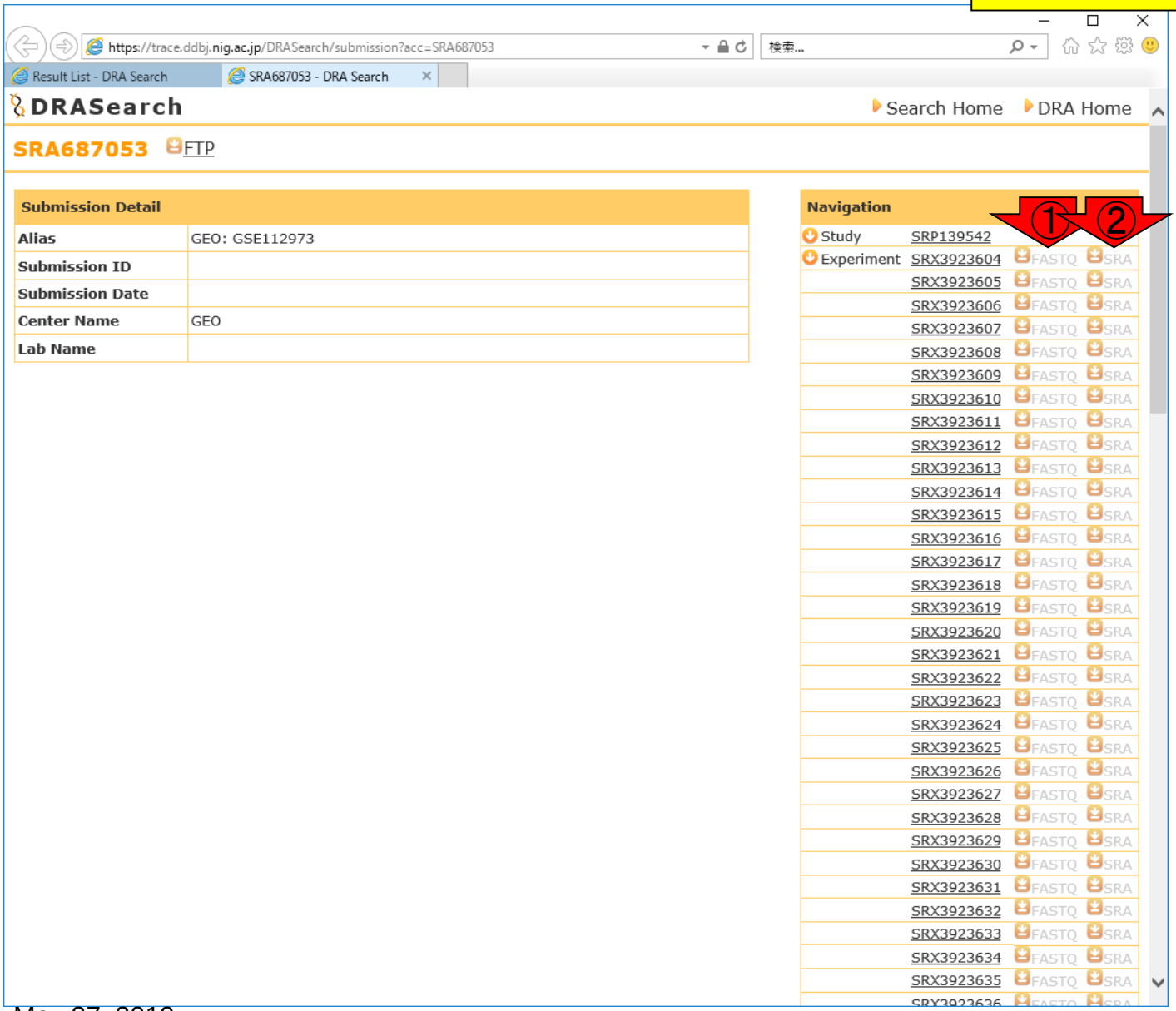
最新のデータだと…

①最後のページに飛んだところ。②SUBMITTEDの日付もないが、数字も大きいのでかなり最近公開されたものなのでしょう。例えば③をクリックすると…



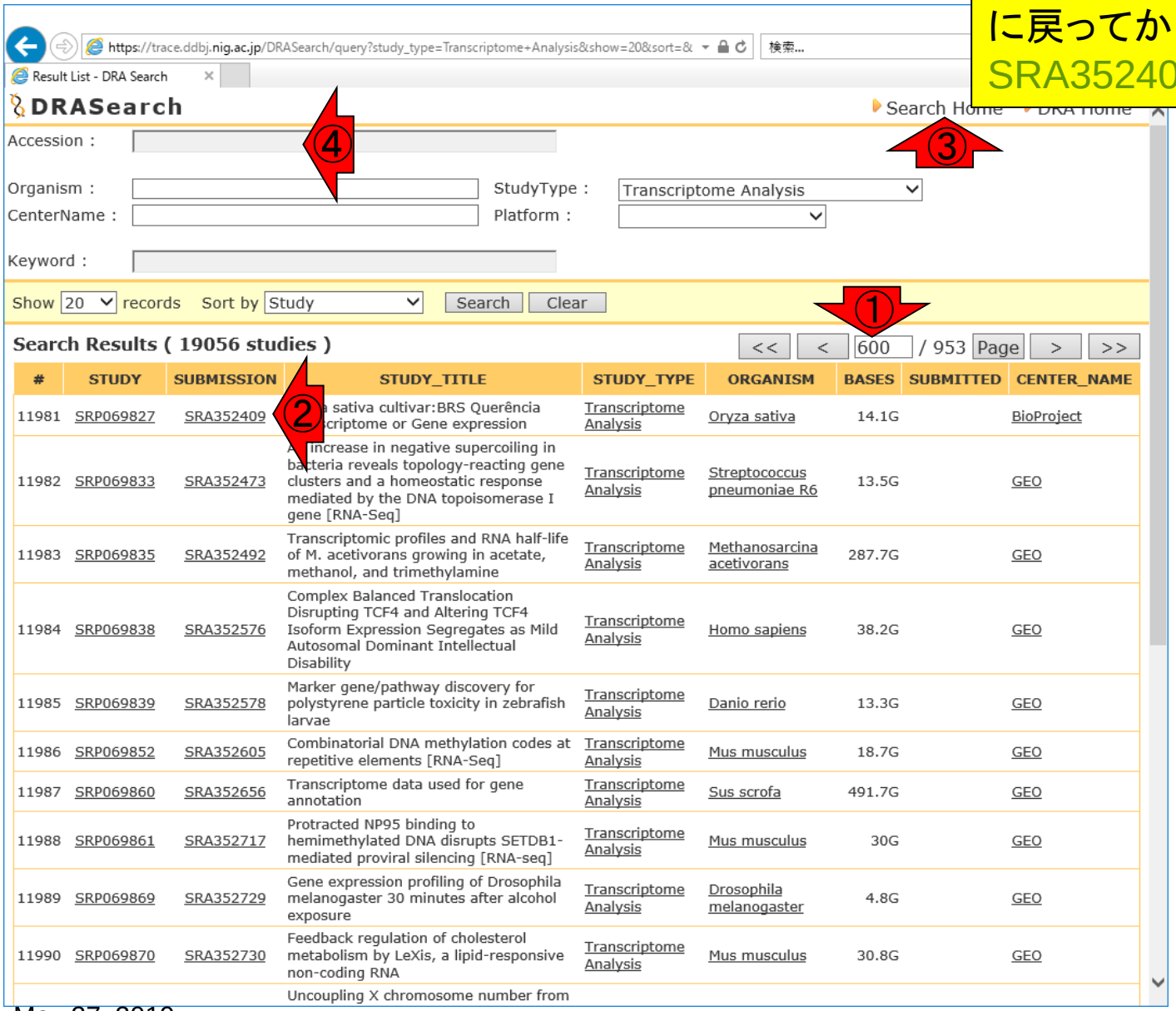
最新のデータだと…

こんな感じになりました。この場合は①FASTQどころか②SRAもまだダウンロードできないようですね。こういうこともあります



そこそこのデータだと…

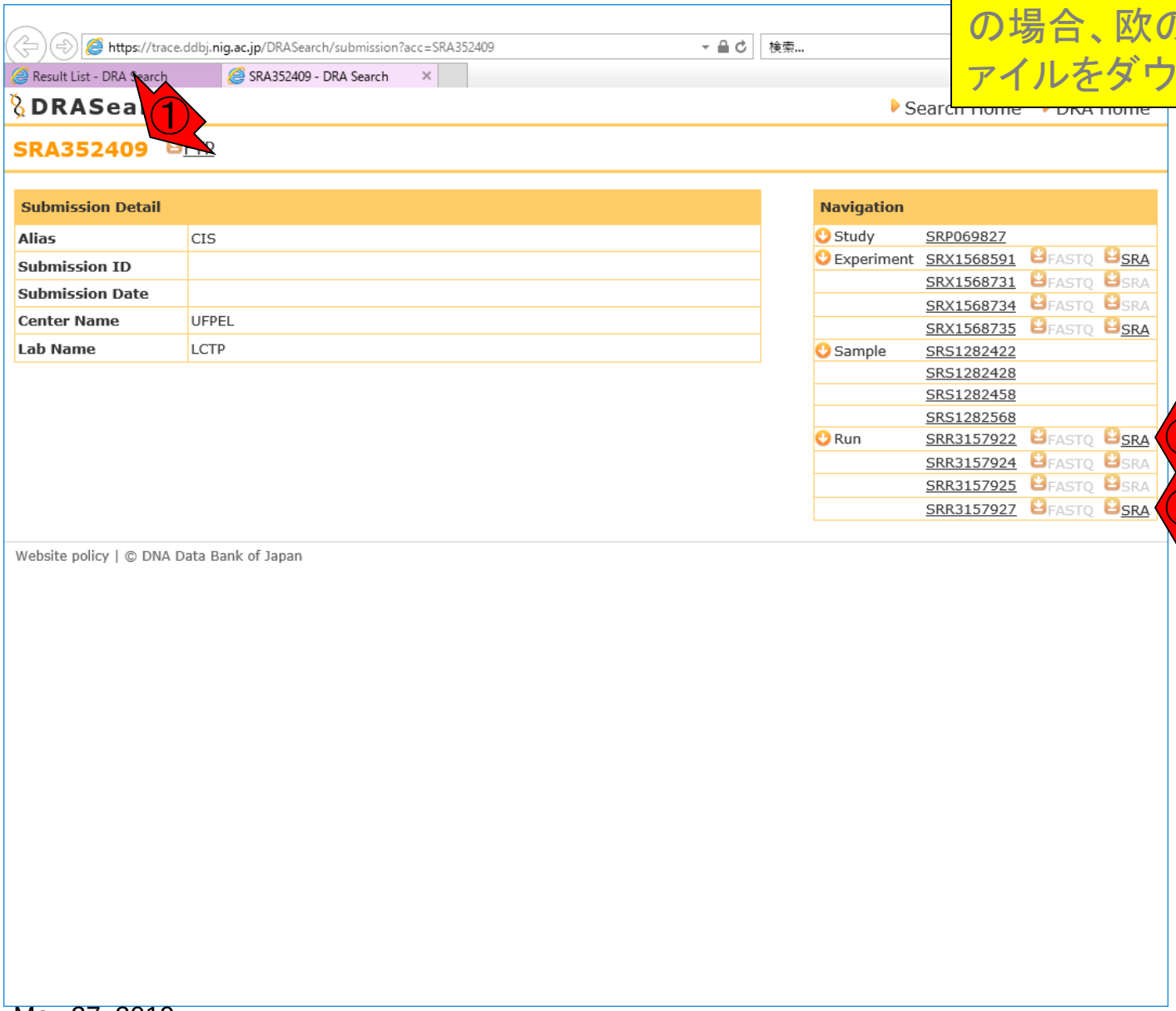
①ここを600とかにして、②SRA352409を見てみる。見る日によっても位置は異なるかもしれないので、③で一旦Search homeに戻ってから、④のAccessionのところにSRA352409と打ち込んでもいいかも…



#	STUDY	SUBMISSION	STUDY_TITLE	STUDY_TYPE	ORGANISM	BASES	SUBMITTED	CENTER_NAME
11981	SRP069827	SRA352409	a sativa cultivar:BRS Querência scriptome or Gene expression	Transcriptome Analysis	Oryza sativa	14.1G		BioProject
11982	SRP069833	SRA352473	A increase in negative supercoiling in bacteria reveals topology-reacting gene clusters and a homeostatic response mediated by the DNA topoisomerase I gene [RNA-Seq]	Transcriptome Analysis	Streptococcus pneumoniae R6	13.5G		GEO
11983	SRP069835	SRA352492	Transcriptomic profiles and RNA half-life of M. acetivorans growing in acetate, methanol, and trimethylamine	Transcriptome Analysis	Methanosarcina acetivorans	287.7G		GEO
11984	SRP069838	SRA352576	Complex Balanced Translocation Disrupting TCF4 and Altering TCF4 Isoform Expression Segregates as Mild Autosomal Dominant Intellectual Disability	Transcriptome Analysis	Homo sapiens	38.2G		GEO
11985	SRP069839	SRA352578	Marker gene/pathway discovery for polystyrene particle toxicity in zebrafish larvae	Transcriptome Analysis	Danio rerio	13.3G		GEO
11986	SRP069852	SRA352605	Combinatorial DNA methylation codes at repetitive elements [RNA-Seq]	Transcriptome Analysis	Mus musculus	18.7G		GEO
11987	SRP069860	SRA352656	Transcriptome data used for gene annotation	Transcriptome Analysis	Sus scrofa	491.7G		GEO
11988	SRP069861	SRA352717	Protracted NP95 binding to hemimethylated DNA disrupts SETDB1-mediated proviral silencing [RNA-seq]	Transcriptome Analysis	Mus musculus	30G		GEO
11989	SRP069869	SRA352729	Gene expression profiling of Drosophila melanogaster 30 minutes after alcohol exposure	Transcriptome Analysis	Drosophila melanogaster	4.8G		GEO
11990	SRP069870	SRA352730	Feedback regulation of cholesterol metabolism by LeXis, a lipid-responsive non-coding RNA	Transcriptome Analysis	Mus musculus	30.8G		GEO
			Uncoupling X chromosome number from					

そこそこのデータだと…

①同じSubmission IDでも、②一部のSRA形式ファイルのみしかダウンロードできないようなものもあります。実は①SRA352409の場合、欧のEMBL-EBI ENAでFASTQファイルをダウンロード可能…



公共DB

NGSデータの公共DBは、①DDBJ SRA、②NCBI SRA、③EMBL-EBI ENAの三極で運用されており、データ共有がなされている。とはいえ、タイムラグは結構あるので注意してください

- イントロ | 一般 | 配列取得 | トランスクリプトーム配列 | GenomicFeatures(Lawrence_20
- イントロ | 一般 | 配列取得 | トランスクリプトーム配列 | biomaRt(Durinck_2009)(last m
- イントロ | 一般 | 読み込み | xlsx形式 | openxlsx(last modified 2015/11/15)
- イントロ | NGS | 様々なプラットフォーム(last modified 2016/03/24)
- イントロ | NGS | qPCRやmicroarrayなどとの比較 (last modified 2014/11/12)
- イントロ | NGS | 可視化(ゲノムブラウザやViewer) (last modified 2016/12/22)
- イントロ | NGS | 配列取得 | FASTQ or SRA | 公共DBから (last modified 2015/02/23)
- イントロ | NGS | 配列取得 | FASTQ or SRA | SRAdb(Zhu_2013) (last modified 2015/02/24)
- イントロ | NGS | 配列取得 | シミュレーションデータ | について (last modified 2015/01/18)
- イントロ | NGS | 配列取得 | シミュレーションデータ | ランダムな塩基配列の生成から (last modified 2015/01/18)
- イントロ | NGS | アノテーション情報取得 | について (last modified 2014/03/26)
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取
- イントロ | NGS | アノテーション情報取

イントロ | NGS | 配列取得 | FASTQ or SRA | 公共DBから NEW

次世代シーケンサ(NGS)から得られる塩基配列データを公共データベースから取得する際には以下を利用します。マイクロアレイデータ取得のときと同様、NGSデータもArrayExpress経由でダウンロードするのがいいかもしれません。メタデータの全貌を把握しやすいこと、生データ(raw data)だけでなく加工済みのデータ(processed data)がある場合にはその存在がすぐにわかることなど、操作性の点で他を凌駕していると思います。上記でも触れているようにFASTQファイルのダウンロードからマッピングまでを行うのはエンドユーザレベルでは大変ですが、submitterが提供してくれている場合は(まだまだ少ないようですが)リファレンス配列へのマップ後のデータ、つまりBAM形式ファイルの提供もすでに始まっているようです。2014年6月26日に知りました(DDBJ児玉さんありがとうございますm(_ _)m)。

データの形式は基本的にSanger typeのFASTQ形式です。FASTA形式はリードあたり二行(idの行と配列の行)で表現します。FASTQ形式はリードあたり4行(@から始まるidの行と配列の行、および+から始まるidの行とbase callの際のqualityの行)で表現します。FASTQ形式は、Sangerのものがデファクトスタンダード(業界標準)です。かつてIlluminaのプラットフォームから得られるのはFASTQ-like formatという表現がなされていたようです(Cock et al., Nucleic Acids Res., 2010)。しかし少なくとも2013年頃には、IlluminaデータもBaseSpaceやCASAVA1.8のconfigureBclToFastq.plなどを用いることで業界標準のFASTQ形式(つまりSanger typeのデータ)に切り替えられるようですし、NCBI SRAなどの公共DBから取得するデータは全てはSanger typeのデータになっていたと思います(Kibukawa E., テクニカルサポートウェビナー, 2013)。

① DDBJ Sequence Read Archive (DRA): Kodama et al., Nucleic Acids Res., 2012
③ EMBL-EBI European Nucleotide Archive (ENA): Leinonen et al., Nucleic Acids Res., 2015
② NCBI Sequence Read Archive (SRA): NCBI Resource Coordinators., Nucleic Acids Res., 2014

ArrayExpress: Kolesnikov et al., Nucleic Acids Res., 2015
- GEO: Barrett et al., Nucleic Acids Res., 2013
- DBCLS SRA: Nakazato et al., PLoS One, 2013