

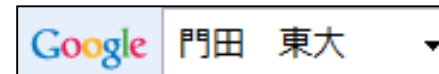
ビッグデータ解析とR

東京大学・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ユニット

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/>



理化学研究所 HPCI 計算生命科学推進プログラム

RIKEN HPCI Program for Computational Life Sciences



産業技術総合研究所 ゲノム情報研究センター HPCI 人材養成プログラム

HPCI Fostering Human Resources Program, Computational Biology Research Center, AIST

<http://www.biomedpharminfo.org/>

第3回生命医薬情報学連合大会2014

大会主催：日本バイオインフォマティクス学会 (JSBI) / 日本オミックス医療学会 / 情報計算化学生物学会 (CBI 学会)



本セッションの宣伝
用ポスターです

スポンサーセッション

生命科学における ビッグデータマイニング ——医療への実践を目指して

【時間】 9:00~10:30 【会場】 萩 (2階)

【プログラム (五十音順)】 座長：江口至洋 副プログラムディレクター

石田貴士 (東京工業大学 大学院情報理工学研究所、助教)

「スーパーコンピュータが実現する大規模メタゲノム機能解析」

門田幸二 (東京大学大学院農学生命科学研究科、特任准教授)

「ビッグデータ解析とR」

新井田厚司 (東京大学医科学研究所ヒトゲノム解析センター、特任助教)

「がんの進化と腫瘍内不均一性を理解するための
ゲノム解析とシミュレーション」

松田秀雄 (大阪大学 大学院情報科学研究科、教授)

「脂肪細胞の遺伝子解析

——白色脂肪細胞の寒冷刺激による褐色化の機構解明——」

チュートリアルセッション

フリーソフトRを用いた ビッグデータ解析：

塩基配列解析を中心に ～ソフトウェアR (ver.3.1.0) と
Bioconductor (ver.2.10) を用いた実習中心の講習会～

【時間】 10:50~12:20 【会場】 小会議室4
(2階)

【講師】 門田幸二 (東京大学大学院農学生命科学研究科、特任准教授)

【受講要件】 必要なソフトウェアやパッケージを
インストール済みのノートPCを持参できる方
(詳細はこちらでご確認ください)
<https://mpc.cbrc.jp/modules/tutorial/biomedpharminfo2014.html>



【要旨】

フリーソフトRは、一般的に統計解析ソフトだというイメージがあるが、ビッグデータ解析の入門としても最適であり、次世代シーケンサーデータ取得やゲノム解析など多様な解析が可能である。本講習会では、特に統計解析以外について、Rで出来る塩基配列解析の全体像の把握を目的としたチュートリアルを行う。Rを使いこなせる楽しさや、バイオインフォマティクスの思考回路の一層に触れる場となれば幸いである。

理化学研究所 HPCI 計算生命科学推進プログラム

RIKEN HPCI Program for Computational Life Sciences



産業技術総合研究所 ゲノム情報研究センター HPCI 人材養成プログラム

HPCI Fostering Human Resources Program, Computational Biology Research Center, AIST

<http://www.scls.riken.jp/>

門田は「生命科学におけるビッグデータマイニング-医療への実践を目指して」の課題に直接取り組んでいるわけではありません。

<http://www.biomedpharminfo.org/>

第3回生命医薬情報学

大会主催：日本バイオインフォマティクス学会 (JSB) / 日本オミックス医療学会



HPCI 戦略プログラム 分野1
「予測する生命科学・医療および創薬基盤」

Japanese | English

[リンク](#) [交通アクセス](#) [お問い合わせ](#)

HOME

SCLSについて

研究内容を知る

人材育成・教育

情報ライブラリ

スパコンを利用する

SCLSを学ぼう

サイト内検索

検索

スポンサーセッション

生命科学における ビッグデータマイニング ——医療への実践を目指して

【時間】 9:00~10:30 【会場】

【プログラム (五十音順)】 座長：江口至洋 副座長：

石田貴士 (東京工業大学 大学院情報理工学系研究科、助教)

「スーパーコンピュータが実現する大規模メ

門田幸二 (東京大学大学院農学生命科学研究科、特任准教授)

「ビッグデータ解析とR」

新井田厚司 (東京大学医科学研究所ヒトゲノム解析センター)

「がんの進化と腫瘍内不均一性を理解するた

ゲノム解析とシミュレーション」

松田秀雄 (大阪大学 大学院情報科学研究科、教授)

「脂肪細胞の遺伝子解析

——白色脂肪細胞の寒冷刺激による褐色化



研究者に聞く

Vol.2 Jan Moren



最新のお知らせ

一覧へ

2014/07/14 [人材育成・教育](#)

2014年度第2期SCLS計算機システム(京コンピュータ
互換スーパーコンピュータシステム) 利用の公募開始
(7/14)

2014/07/11 [メディア・リリース](#)

マルチスケール・マルチフィジックス心臓シミュレ
ータ UT-Heart ④(映像コンテンツ7/11公開)

2014/07/01 [メディア・リリース](#)

TBSテレビ「夢の扉+」「スパコン」京で未来を予測せ
よ! ④(7/6 18:30-19:00放送予定)、分野1取材協力

2014/06/30 [セミナー・シンポジウム](#)

だれにでもわかる拡張サンプリングシミュレーション

4つの研究課題

課題1 細胞内分子ダイナミクス



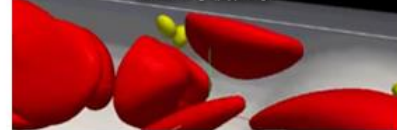
細胞内分子ダイナミクスのシミュレーション

課題2 創薬応用



創薬応用シミュレーション

課題3 予測医療



予測医療に向けた階層統合シミュレーション

課題4 大規模生命データ解析



大規模生命データ解析

Oct 04 2014, HPCI workshop

理化学研究所 HPCI 計算生命科学推進プログラム

RIKEN HPCI Program for Computational Life Sciences

産業技術総合研究所 ゲノム情報 HPCI 人材養成プログラム

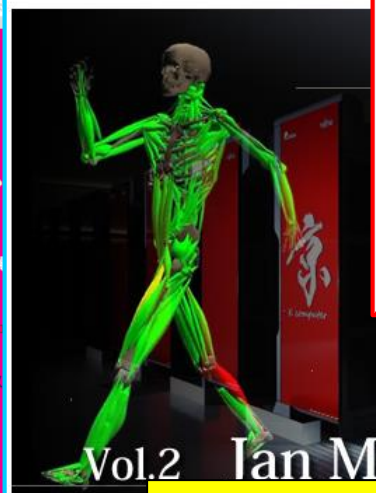
HPCI Fostering Human Resources Program, Computational Biology R

http://www.biomedpharminfo.org/
第3回生命医薬情報学
大会主催：日本バイオインフォマティクス学会 (JSBI) 日本オミックス医療学会



スポンサーセッション
生命科学における
ビッグデータマイニ
ング—医療への実践を目指して
【時間】 9:00~10:30 【会場】

【プログラム (五十音順)】
石田貴士 (東京工業大学 大学院情報理工学研究所、助教)
「スーパーコンピュータが実現する大規模メ
門田幸二 (東京大学大学院農学生命科学研究科、特任准教授)
「ビッグデータ解析とR」
新井田厚司 (東京大学医科学研究所ヒトゲノム解析センター)
「がんの進化と腫瘍内不均一性を理解するた
ゲノム解析とシミュレーション」
松田秀雄 (大阪大学 大学院情報科学研究科、教授)
「脂肪細胞の遺伝子解析
—白色脂肪細胞の寒冷刺激による褐色化



4つの研究課題

課題1 創薬

細胞内分子ダイナミクスのシミュレーション

創薬応用シミュレーション

課題4 大規模生命データ解析

代表：宮野 悟 東京大学 医科学研究所

1) 大規模データ解析によるがんのシステム
異常の網羅的解析とその応用

新井田厚司 先生

宮野 悟
(東京大学 医科学研究所)

2) 大規模生体分子ネットワーク解析による
脂肪細胞組織の刺激応答の網羅的解析と
その応用

松田秀雄 先生

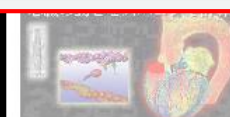
松田 秀雄
(大阪大学大学院情報科学研究科)

3) 次世代シーケンサーデータ解析のための
情報処理システムの開発

石田貴士 先生

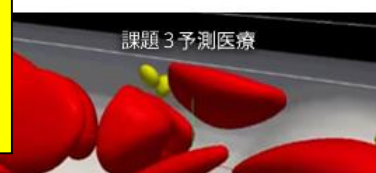
秋山 泰
(東京工業大学大学院情報理工学研究所)

「生命科学におけるビッグデータ
マイニング—医療への実践を目
指して」の課題に直接取り組んで
いるがおそらく私以外の先生方。



2014/07/01 メディア・リリース
TBSテレビ「夢の扉+」「スバコン」京で未来を予測せよ」(7/6 18:30-19:00放送予定)、分野1取材協力

2014/06/30 セミナー・シンポジウム
だれにでもわかる拡張サンプリングシミュレーション



予測医療に向けた階層統合シミュレーション



大規模生命データ解析

理化学研究所 HPCI 計算生命科学推進プログラム

RIKEN HPCI Program for Computational Life Sciences

産業技術総合研究所 ゲノム情報 HPCI 人材養成プログラム

HPCI Fostering Human Resources Program, Computational Biology R

第3回生命医薬情報学連合大会2014

大会主催：日本バイオインフォマティクス学会 (JSBI) / 日本オミックス医療学会 / 情報計算化学生化学会 (CBI学会)

スポンサーセッション

生命科学における ビッグデータマイニング —医療への実践を目指して—

【時間】 9:00~10:30 【会場】 萩 (2階)

【プログラム (五十音順)】

- 石田貴士 (東京工業大学 大学院情報理工学研究所、助教)
「スーパーコンピュータが実現する大規模メタゲノム機能解析」
- 門田幸二 (東京大学 大学院農学生命科学研究科、特任准教授)
「ビッグデータ解析とR」
- 新井田厚司 (東京大学 医科学研究所ヒトゲノム解析センター、特任助教)
「がんの進化と腫瘍内不均一性を理解するための
ゲノム解析とシミュレーション」
- 松田秀雄 (大阪大学 大学院情報科学研究科、教授)
「脂肪細胞の遺伝子解析
—白色脂肪細胞の寒冷刺激による褐色化の機構解明—」

チュートリアル

フリー ビッグ 塩基配列解

【時間】 10:50

【講師】 門田幸二 (東京大学大学院農学生命科学研究科、特任准教授)

【受講要件】 必要なソフトウェアやパッケージをインストール済みのノートPCを持参できる方
(詳細はこちらでご確認ください)
<http://mpci.cbrc.jp/modules/tutorial/biomedpharminfo2014.html>

【要旨】

課題4 大規模生命データ解析

代表：宮野 悟 東京大学 医科学研究所

1) 大規模データ解析によるがんのシステム
異常の網羅的解析とその応用

新井田厚司 先生

宮野 悟
(東京大学 医科学研究所)

2) 大規模生体分子ネットワーク解析による
脂肪細胞組織の刺激応答の網羅的解析と
その応用

松田秀雄 先生

松田 秀雄
(大阪大学大学院情報科学研究科)

3) 次世代シーケンサーデータ解析のための
情報処理システムの開発

石田貴士 先生

秋山 泰
(東京工業大学大学院情報理工学研究所)

門田は前座です



「生命科学におけるビッグデータ
マイニング -医療への実践を目
指して」の課題に直接取り組んで
いるがおそらく私以外の先生方。

自己紹介



- 1974年高知県生まれ
- 2002年3月
 - 東京大学・大学院農学生命科学研究科・応用生命工学専攻 博士課程修了
 - 学位論文:「cDNAマイクロアレイを用いた遺伝子発現解析手法の開発」
(指導教官:清水謙多郎教授)
- 2002/4/1~
 - 産総研・生命情報科学研究センター(CBRC) 産総研特別研究員
 - マイクロアレイ解析手法開発
- 2003/11/1~
 - 放医研・先端遺伝子発現研究センター 研究員
 - 一次元電気泳動波形解析手法開発
- 2005/2/16~
 - 東京大学・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プログラム
 - マイクロアレイ解析手法開発
 - RNA-seqデータ解析手法開発

研究は(トランスクリプトーム解析周辺の)手法開発系ですが、最近はフリーソフトウェアR関連のハンズオンセミナーなど教育(人材養成)がメイン。





東京大学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ユニット

Agricultural Bioinformatics Research Unit

+ サイトマップ + English

受講生の方へ 研究者の方へ

- + ホーム
- + 本ユニットについて
- + メンバー
- + 教育プログラム
- + 研究フォーラム
- + イベント
- + お問い合わせ
- + リンク
- + モバイルサイト

ホーム > 教育プログラム > 各講義のページ



各講義のページ

(科目名をクリックすると各講義のページに移動します)

先端トピックス
セミナー・討論形式
研究指導

農学生命情報科学特別演習

農学生命情報科学特論 I **農学生命情報科学特論 II** **農学生命情報科学特論 III** **農学生命情報科学特論 IV**

方法論
講義・実習を一体化

生物配列統計学 システム生物学概論 知識情報処理
オーム情報解析 機能ゲノム学 分子モデリングと分子シミュレーション

基礎
講義・実習を一体化

ゲノム情報解析基礎 構造バイオインフォマティクス基礎
生物配列解析基礎 バイオスタティスティクス基礎論

カテゴリー	科目名	学期・単位	実施曜日
基礎	1. 生物配列解析基礎	夏・1	火曜
	生命科学のためのデータベースの利用と基本的な解析手法について講義します。データベースの基礎、配列データベース、機能データベース、ホモロジー検索、モチーフ解析などの基本的な手法について解説します。		

講義風景(平成26年度)



アグリバイオインフォマティクスでは、主にRを用いて100人規模の実践的なハンズオン大学院講義を行っています。



(Rで)マイクロアレイデータ解析

(last modified 2014/05/17, since 2005)

What's new?

- ・門田幸二 著 [シリーズ Useful R 第...](#) 最近の知見や、ROKU法 (Kadota) 書籍中のマイクロアレイ解析 [...] に掲載してあります。(2014/04/17)
- ・お知らせは主に [\(Rで\)塩基配列解析](#) 講演資料なども [\(Rで\)塩基配列解析](#)

- ・ [はじめに](#) (last modified 2014/05/17)
- ・ [過去のお知らせ](#) (last modified 2014/06/30) **NEW**
- ・ [Rのインストールと起動](#) (last modified 2014/05/14)
- ・ [Rの昔のバージョンのインストール](#)
- ・ [使用例\(初心者向け\)](#) (last modified 2014/06/21) **NEW**
- ・ [サンプルデータ](#) (last modified 2014/06/21) **NEW**
- ・ [書籍 11 について](#) (last modified 2014/05/12)
- ・ 書籍 [トランスクリプトーム解析](#)
- ・ 書籍 [トランスクリプトーム解析](#)
- ・ 書籍 [トランスクリプトーム解析](#)
- ・ 書籍 [トランスクリプトーム解析](#)
- ・ 書籍 [トランスクリプトーム解析](#)
- ・ 書籍 [トランスクリプトーム解析](#)
- ・ 書籍 [トランスクリプトーム解析](#)

私の講義は、この2つのウェブページを利用しています。

(Rで)塩基配列解析

～NGS、RNA-seq、ゲノム、トランスクリプトーム、正規化、発現変動、統計、モデル、バイオインフォマティクス～
(last modified 2014/07/07, since 2010)

What's new?

- ・ このウェブページはフリーソフト R と利用可能なパッケージの多くをインストール済みである前提で記述していますので、[Rのインストールと起動](#) を参考に必要なパッケージのインストールを行ってください。
- ・ 2014年7月22日に [イルミナウェビナー](#) で話します。興味ある方はどうぞ。(2014/06/30) **NEW**
- ・ 門田幸二 著 [シリーズ Useful R 第7巻トランスクリプトーム解析](#) 刊行 (共立出版)
- ・ [マップ後 | 配列長とカウント 数の関係](#) のところで、boxplotでの描画の際に param 個で分割 (20 分割など) するテクニックとして `floor(nrow(data)/param)+1` としていましたが、これだと割り切れる場合でも +1 してしまうことが判明したため `ceiling(nrow(data)/param)` に修正しました (佐伯亘平氏提供情報)。(2014/07/03) **NEW**
- ・ 2014年9月1日～12日に「[バイオインフォマティクス人材育成カリキュラム\(次世代シーケンサ\)速習コース](#)」を東大農で開催します。受講申込は6/24夕方に締め切りでしたが、TA申込み枠はまだ若干余裕があります。TA申込みが全日程受講申込締め切り後の6/24から7/3朝までできない状態になっていたようで失礼しました m(_ _)m。7/4の10:00ごろに復旧しております。(2014/07/04) **NEW**
- ・ [参考資料\(講義、講習会、本など\)](#) の項目を追加しました。(2014/07/03) **NEW**

- ・ [はじめに](#) (last modified 2014/01/30)
- ・ [参考資料\(講義、講習会、本など\)](#) (last modified 2014/07/07) **NEW**
- ・ [過去のお知らせ](#) (last modified 2014/06/30) **NEW**
- ・ [Rのインストールと起動](#) (last modified 2014/05/14)
- ・ [サンプルデータ](#) (last modified 2014/06/21) **NEW**
- ・ [書籍 | トランスクリプトームについて](#) (last modified 2014/05/12)

本日の講演資料はここからダウンロードできます

[トップページへ](#)

(Rで)マイクロアレイデータ解析

(last modified 2014/05/17, since 2005)

What's new?

- ・門田幸二 著 [シリーズ Useful R 第7巻](#) 最近の知見や、ROKU法 (Kadota) による書籍中のマイクロアレイ解析 [...] に掲載してあります。(2014/07/07)
- ・お知らせは主に [\(Rで\)塩基配列解析](#) 講演資料なども [\(Rで\)塩基配列解析](#)

- ・ [はじめに](#) (last modified 2014/05/17)
- ・ [過去のお知らせ](#) (last modified 2014/06/30) **NEW**
- ・ [Rのインストールと起動](#) (last modified 2014/05/14)
- ・ [Rの昔のバージョンのインストール](#)
- ・ [使用例\(初心者向け\)](#) (last modified 2014/06/21) **NEW**
- ・ [サンプルデータ](#) (last modified 2014/06/21) **NEW**
- ・ [書籍 11 について](#) (last modified 2014/05/12)
- ・ 書籍 | [トランスクリプトーム解析](#) |
- ・ 書籍 | [トランスクリプトーム解析](#) |
- ・ 書籍 | [トランスクリプトーム解析](#) |
- ・ 書籍 | [トランスクリプトーム解析](#) |
- ・ 書籍 | [トランスクリプトーム解析](#) |
- ・ 書籍 | [トランスクリプトーム解析](#) |
- ・ 書籍 | [トランスクリプトーム解析](#) |
- ・ 書籍 | [トランスクリプトーム解析](#) |

ウェブページの内容をまとめたものが2014年4月出版

(Rで)塩基配列解析

～NGS、RNA-seq、ゲノム、トランスクリプトーム、正規化、発現変動、統計、モデル
(last modified 2014/07/07, since 2010)

What's new?

- ・ このウェブページはフリーソフト R と利用可能なパッケージの多くをインストール済なので、[Rのインストールと起動](#) を参考に必要なパッケージのインストールを
- ・ 2014年7月22日に [イルミナウェビナー](#) で話します。興味ある方はどうぞ。(2014/06/30)
- ・ 門田幸二 著 [シリーズ Useful R 第7巻トランスクリプトーム解析](#) 刊行 (共立出版)
- ・ [マップ後 | 配列長とカウント 数の関係](#) のところで、boxplotでの描画の際に param 値のニックとして `floor(nrow(data)/param)+1` としていましたが、これだと割り切れる場合のために `ceiling(nrow(data)/param)` に修正しました (佐伯亘平氏提供情報)。(2014/07/04) **NEW**
- ・ 2014年9月1日～12日に「[バイオインフォマティクス人材育成カリキュラム\(次世代\)](#)」東大農で開催します。受講申込は6/24夕方に締め切りでしたが、TA申込み枠はTA申込みが全日程受講申込締め切り後の6/24から7/3朝までできない状態に `m(_).m`。7/4の10:00ごろに復旧しております。(2014/07/04) **NEW**
- ・ [参考資料\(講義、講習会、本など\)](#) の項目を追加しました。(2014/07/03) **NEW**

- ・ [はじめに](#) (last modified 2014/01/30)
- ・ [参考資料\(講義、講習会、本など\)](#) (last modified 2014/07/07) **NEW**
- ・ [過去のお知らせ](#) (last modified 2014/06/30) **NEW**
- ・ [Rのインストールと起動](#) (last modified 2014/05/14)
- ・ [サンプルデータ](#) (last modified 2014/06/21) **NEW**
- ・ [書籍 | トランスクリプトームについて](#) (last modified 2014/05/12)



本日の講演資料はここからダウンロードできます

[トップページへ](#)

約 92,200 件 (0.12 秒)

(Rで)塩基配列解析

www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html ▼

(2014/06/25); 門田幸二 著シリーズ Useful R 第7巻トランスクリプトーム解析刊行(共立出版) **NGS**などから得られた短い塩基配列(**short read**)データ解析をRで行うための一連の手続きをまとめているものであり、特にアグリバイオインフォマティクス教育 ...

次世代シーケンサ(NGS)解析で使われるソフトの簡単なまとめ ...

d.hatena.ne.jp/sesejun/20100521/p1 ▼

2010/05/21 - 次世代シーケンサ(NGS)解析で使われるソフトの簡単なまとめ.

2010/5/25. ... 既に解析をガシガシやられている先人の方から、こんなソフト使ってるよーとか、そんなん使わん、とか突っ込み歓迎です。あとソフトが見つからないところがある ...

次世代シーケンサー お悩みカウンセリング

www.mss.co.jp/businessfield/bioinformatics/ngs/ ▼

参照配列(既知のゲノム配列)にリードをマッピング(参照配列のどこに該当するのかを決める)する一次解析と、解析手法ごとの二次解析の二段階の処理が必要になります。マッピングが必要な解析. **NGS**機器からの配列データ(FASTQ)詳細・参照配列詳細.

NGS Surfer's Wiki | ツール情報 - Cell Innovation Program

<https://cell-innovation.nig.ac.jp/wiki/tiki-index.php?page=ツール情報> ▼

2014/03/25 - sesejun先生のブログ Loud Minority(次世代シーケンサ(NGS)解析で使われるソフトの簡単なまとめ) (external link) ... マッパーの情報はありますが、エビジェネティック関連のツールのカテゴリが無いので備忘的に。高次解析用 ...

よく分かる次世代シーケンサー解析を開催します。3月17日 ...

Rの内容中心ですがNGS関連キーワードでもよく引っかけられます。Rは京にもインストールされており、ビッグデータの前処理や後処理で使われているようです。





産業技術総合研究所 ゲノム情報研究センター
HPCI 人材養成プログラム
HPCI Fostering Human Resources Program, Computational Biology Research Center, AIST

門田はHPCI人材養成プログラムの一部を例年担当
させていただきます

<http://www.biomedpharminfo.org/>

第3回生命医薬情報学連合
大会主催：日本バイオインフォマティクス学会 (JSBI) / 日本オミックス医療学会 / 情報計算化学

HPCI 戦略プログラム 分野 1

「予測する生命科学・医療および創薬基盤」

代表機関(独) 理化学研究所

人材養成プログラム

実施機関(独) 産業技術総合研究所 ゲノム情報研究センター

スポンサーセッション

生命科学における
ビッグデータマイニング
——医療への実践を目指して

【時間】 9:00~10:30 【会場】 萩

【プログラム(五十音順)】 座長：江口至洋 副プログラムディレクター

石田貴士 (東京工業大学 大学院情報理工学系研究科、助教)

「スーパーコンピュータが実現する大規模メタゲノム

門田幸二 (東京大学大学院農学生命科学研究科、特任准教授)

「ビッグデータ解析とR」

新井田厚司 (東京大学医科学研究所ヒトゲノム解析センター、特任助教)

「がんの進化と腫瘍内不均一性を理解するための

ゲノム解析とシミュレーション」

松田秀雄 (大阪大学 大学院情報科学研究科、教授)

「脂肪細胞の遺伝子解析

——白色脂肪細胞の寒冷刺激による褐色化の機構解



top

outline

seminar

workshop

tutorial

e-learning

links

contact

e-learning

～平成26年度 開講しました～ H25年 HPCIセミナー 新規コンテンツ追加
どなたでもお好きな時間にインターネット経由で学べるeラーニング
バイオインフォ・創薬インフォ・最先端の研究紹介セミナー等を収録しています。

詳細は
こちら

tutorial

～平成26年度 講習会ご案内～ 1人1台のPCを用いた実習付き講習会を開催
創薬インフォマティクス実習コース 受講申込受付中です。
また創薬コース@大阪 キャンセル待ち、特論のみ聴講申込受付中です。

詳細は
こちら

seminar

～平成26年度 日程・講師確定～ 10月以降金曜全12回、講演タイトル公開中
最先端の研究セミナーを一般公開、どなたでも聴講頂けます。
平成23, 24, 25年度の一部のHPCIセミナーはe-learning化されています。

詳細は
こちら

workshop

～平成26年度 開催日時・講演者決定～ 10月4日(土)9時～
HPC(High Performance Computing)を用いた生命科学の最先端の話題に
フォーカスしたワークショップを生命医薬情報学連合大会2014において開催予定

詳細は
こちら



<http://www.biomedpharminfo.org/>

第3回生命医薬情報学連合大会2014
大会主催：日本バイオインフォマティクス学会 (JSBI) / 日本オミックス医療学会 / 情報計算生命科学学会 (CBI 学会)

スポンサーセッション

生命科学における
ビッグデータマイニング
——医療への実践を目指して

【時間】 9:00~10:30 【会場】 萩 (2階)

【プログラム (五十音順)】 座長：江口至洋 副プログラムディレクター

石田貴士 (東京工業大学 大学院情報理工学研究所、助教)
「スーパーコンピュータが実現する大規模メタゲノム機能解析」

門田幸二 (東京大学大学院農学生命科学研究所、特任准教授)
「ビッグデータ解析とR」

新井田厚司 (東京大学医科学研究所ヒトゲノム解析センター、特任助教)
「がんの進化と腫瘍内不均一性を理解するための
ゲノム解析とシミュレーション」

松田秀雄 (大阪大学 大学院情報科学研究科、教授)
「脂肪細胞の遺伝子解析
——白色脂肪細胞の寒冷刺激による褐色化の機構解明——」

チュートリアル

フリー
ビッグ
塩基配列

【時間】 10:

【講師】 門田幸二

【受講要件】 必須
イン
PC
MS

【要旨】
フリーソフトRは
データ解析の入門とし
て多様な解析が可能
な塩基配列解析の全体
像や、バイオイン

・ バイオインフォマティクス実習コース

- バイオインフォマティクスの基礎知識・実践技術を短期間に習得 -

第一線の研究者が講師として、バイオ情報に指導します。計算機実習は1人1台のPC混ぜたカリキュラムで、バイオインフォマで講義・実習を行い、受講を希望するものる知識や技術を設定した受講要件がありまれまで当センターで実施してきた生命情報コースII、生命情報科学人材養成コンソーシアム バイオインフォマティクス実習コースのカリキュラムを元に企画し直しました。

HPCI人材養成プログラムの一環として、Linuxに比べて敷居の低いRを用いた(ビッグ)データ解析のノウハウを伝授しています。

大量の配列データも怖くない!! Windows上のLinux環境で高速・簡単に配列解析

A-1	Linux, Perl基礎	2014年10月28日(火)
A-2	配列解析	2014年10月29-30日(水、木)

効率的なWindows上のLinux環境で次世代シーケンサーからのデータを解析

A-3	ChIP-seqデータ解析およびENCODEプロジェクトなどによる既存のデータの活用	2014年11月5-6日(水、木)
-----	--	-------------------

フリーウェア 統計解析パッケージ R を使った NGS データ解析を基礎から学ぶ

B-1	R基礎	2015年3月4日(水)
B-2	Rでゲノム・トランスクリプトーム解析:CpG解析から機能解析まで	2015年3月5-6日(木、金)
B-3	多変量データ解析/遺伝子ネットワーク解析	2015年3月12-13日(木、金)

産業技術総合研究所 ゲノム情報研究センター
HPCI 人材養成プログラム
HPCI Fostering Human Resources Program, Computational Biology Research Center, AIST

<http://www.biomedpharminfo.org/>

第3回生命医薬情報学連合大会2014

大会主催：日本バイオインフォマティクス学会 (JSBI) / 日本オミックス医療学会 / 情報計算化学生物学会 (CBI学会)



スポンサーセッション

生命科学における
ビッグデータマイニング
——医療への実践を目指して

【時間】 9:00~10:30 【会場】 萩 (2階)

【プログラム (五十音順)】 座長：江口至洋 副プログラムディレクター

石田貴士 (東京工業大学 大学院情報理工学研究所、助教)
「スーパーコンピュータが実現する大規模メタゲノム機能解析」

門田幸二 (東京大学大学院農学生命科学研究科、特任准教授)
「ビッグデータ解析とR」

新井田厚司 (東京大学医科学研究所ヒトゲノム解析センター、特任助教)
「がんの進化と腫瘍内不均一性を理解するための
ゲノム解析とシミュレーション」

松田秀雄 (大阪大学 大学院情報科学研究科、教授)
「脂肪細胞の遺伝子解析
——白色脂肪細胞の寒冷刺激による褐色化の機構解明——」

チュートリアルセッション

フリーソフトRを用いた
ビッグデータ解析：
塩基配列解析を中心に

～ソフトウェアR (ver.3.1.0) と
Bioconductor (ver.2.10) を用いた実習中心の講習会～

【時間】 10:50~12:20 【会場】 小会議室4 (2階)

【講師】 門田幸二 (東京大学大学院農学生命科学研究科、特任准教授)

【受講要件】 必要なソフトウェアやパッケージを
インストール済みのノートPCを持参できる方



【要旨】

フリーソフトRは
データ解析の入門とし
て多様な解析が可能
な塩基配列解析の全体
像や、バイオイン

フリーウェア 統計解析パッケージ R を使った NGS データ解析を基礎から学ぶ

B-1	R基礎	2015年3月4日(水)
B-2	Rでゲノム・トランスクリプトーム解析：CpG解析から 機能解析まで	2015年3月5-6日(木、金)
B-3	多変量データ解析／遺伝子ネットワーク解析	2015年3月12-13日(木、金)

HPCI人材養成プログラムの一環として、Linuxに比べて敷居の低いRを用いた(ビッグ)データ解析のノウハウを伝授しています。受講者の要望を踏まえ、今年度は講習時間を倍増(丸1日→2日)。

産業技術総合研究所 ゲノム情報研究センター
HPCI 人材養成プログラム
HPCI Fostering Human Resources Program, Computational Biology Research Center, AIST

<http://www.biomedpharminfo.org/>
第3回生命医薬情報学連合大会2014
大会主催：日本バイオインフォマティクス学会 (JSBI) / 日本オミックス医療学会 / 情報計算化学生物学会 (CBI 学会)



HPCI人材養成プログラムの一環として、Linuxに比べて敷居の低いRを用いた(ビッグ)データ解析のノウハウを伝授しています。受講者の要望を踏まえ、今年度は講習時間を倍増(丸1日→2日)。メインは時間数的にも来年3月のほうですが、プログラム内部の説明、マイクロアレイ解析希望者への対応、こんな感じでやってます的な話が本日10:50-12:20のセッション。

スポンサーセッション

生命科学における
ビッグデータマイニング
——医療への実践を目指して

【時間】 9:00~10:30 【会場】 萩 (2階)

【プログラム (五十音順)】 座長：江口至洋 副プログラムディレクター

石田貴士 (東京工業大学 大学院情報理工学研究所、助教)
「スーパーコンピュータが実現する大規模メタゲノム機能解析」

門田幸二 (東京大学大学院農学生命科学研究科、特任准教授)
「ビッグデータ解析とR」

新井田厚司 (東京大学医科学研究所ヒトゲノム解析センター、特任助教)
「がんの進化と腫瘍内不均一性を理解するための
ゲノム解析とシミュレーション」

松田秀雄 (大阪大学 大学院情報科学研究科、教授)
「脂肪細胞の遺伝子解析
——白色脂肪細胞の寒冷刺激による褐色化の機構解明——」

チュートリアルセッション

フリーソフトRを用いた
ビッグデータ解析：
塩基配列解析を中心に

【時間】 10:50~12:20 【会場】 小会議室4 (2階)

【講師】 門田幸二 (東京大学大学院農学生命科学研究科、特任准教授)

【受講要件】 必要なソフトウェアやパッケージを
インストール済みのノートPCを持参できる方
(詳細はこちらでご確認ください)
<https://hpci.cbrc.jp/modules/tutorial/biomedpharminfo2014.html>

【要旨】
フリーソフトRは、一般的に統計解析ソフトだというイメージがあるが、ビッグデータ解析の入門としても最適であり、次世代シーケンサーデータ取得やゲノム解析など多様な解析が可能である。本講習会では、特に統計解析以外について、Rで出来る塩基配列解析の全体像の把握を目的としたチュートリアルを行う。Rを使いこなせる楽しさや、バイオインフォマティクスの思考回路の一端に触れる場となれば幸いである。



Rを使った NGS データ解析を基礎から学ぶ

2015年3月4日(水)

2015年3月5-6日(木、金)

2015年3月12-13日(木、金)

B-2 Rでゲノム・トランスクリプトーム解析：CpG解析から
機能解析まで

B-3 多変量データ解析／遺伝子ネットワーク解析

NGSデータ解析とR

(Rで)塩基配列解析

～NGS、RNA-seq、ゲノム、トランスクリプトーム、正規化、発現変動、統計、モデル、バイオインフォマティクス～
(last modified 2014/07/14, since 2010)

What	•	イントロ		一般		ランダムに行を抽出 (last modified 2013/10/10)
	•	イントロ		一般		任意の文字列を行の最初に挿入 (last modified 2013/10/10)
• この	•	イントロ		一般		任意のキーワードを含む行を抽出(基礎) (last modified 2014/04/11)
すの	•	イントロ		一般		ランダムな塩基配列を生成 (last modified 2014/06/16) NEW
• 201	•	イントロ		一般		任意の長さの可能な全ての塩基配列を作成 (last modified 2013/06/14)
ン	•	イントロ		一般		任意の位置の塩基を置換 (last modified 2013/09/12)
• 201	•	イントロ		一般		指定した範囲の配列を取得 (last modified 2014/03/08)
• 門	•	イントロ		一般		指定したID(染色体やdescription)の配列を取得 (last modified 2014/03/10)
	•	イントロ		一般		翻訳配列(translate)を取得 (last modified 2013/06/14)
• マッ	•	イントロ		一般		相補鎖(complement)を取得 (last modified 2013/06/14)
ニッ	•	イントロ		一般		逆相補鎖(reverse complement)を取得 (last modified 2013/06/14)
した	•	イントロ		一般		逆鎖(reverse)を取得 (last modified 2013/06/14)
• 201	•	イントロ		一般		2連続塩基の出現頻度情報を取得 (last modified 2014/04/14)
東	•	イントロ		一般		3連続塩基の出現頻度情報を取得 (last modified 2013/06/14)
ア	•	イントロ		一般		任意の長さの連続塩基の出現頻度情報を取得 (last modified 2013/06/14)
申入	•	イントロ		一般		Tips 任意の拡張子でファイルを保存 (last modified 2013/09/26)
ム	•	イントロ		一般		Tips 拡張子は同じで任意の文字を追加して保存 (last modified 2013/09/26)
• 参						

配列の切り出しなど基本的な塩基配列解析の多くをカバー。EMBOSS的なイメージ。



NGSデータ解析とR

(Rで)塩基配列解析

～NGS、RNA-seq、ゲノム、トランスクリプトーム、正規化、発現変動、統計、モデル、バイオインフォマティクス～
(last modified 2014/07/14, since 2010)

What	•	イントロ		一般		ランダムに行を抽出 (last modified 2013/10/10)
	•	イントロ		一般		任意の文字列を行の最初に挿入 (last modified 2013/10/10)
• この	•	イントロ		一般		任意のキーワードを含む行を抽出(基礎) (last modified 2014/04/11)
すの	•	イントロ		一般		ランダムな塩基配列を生成 (last modified 2014/06/16) NEW
• 201	•	イントロ		一般		任意の長さの可能な全ての塩基配列を作成 (last modified 2013/06/14)
ン	•	イントロ		一般		任意の位置の塩基を置換 (last modified 2013/09/12)
• 201	•	イントロ		一般		指定した範囲の配列を取得 (last modified 2014/03/08)
• 門	•	イントロ		一般		指定したID(染色体やdescription)の配列を取得 (last modified 2014/03/10)
	•	イントロ		一般		翻訳配列(translate)を取得 (last modified 2013/06/14)
• マッ	•	イントロ		一般		相補鎖(complement)を取得 (last modified 2013/06/14)
ニッ	•	イントロ		一般		逆相補鎖(reverse complement)を取得 (last modified 2013/06/14)
した	•	イントロ		一般		逆鎖(reverse)を取得 (last modified 2013/06/14)
• 201	•	イントロ		一般		2連続塩基の出現頻度情報を取得 (last modified 2014/04/14)
東	•	イントロ		一般		3連続塩基の出現頻度情報を取得 (last modified 2013/06/14)
ア	•	イントロ		一般		任意の長さの連続塩基の出現頻度情報を取得 (last modified 2013/06/14)
申入	•	イントロ		一般		Tips 任意の拡張子でファイルを保存 (last modified 2013/09/26)
ム	•	イントロ		一般		Tips 拡張子は同じで任意の文字を追加して保存 (last modified 2013/09/26)
• 参						

連続塩基の出現頻度解析が可能。ヒトゲノム配列を読み込んでCGの連続塩基が期待値よりも低いことを確認可能(CpG解析)。k-mer解析の基本形に相当。



NGSデータ解析とR

(Rで)塩基配列解析

～NGS、RNA-seq、ゲノム、トランスクリプトーム、正規化、発現変動、統計、モデル、バイオインフォマティクス～
(last modified 2014/07/14, since 2010)

What	•	イントロ		一般		配列取得		ゲノム配列		公共DBから (last modified 2014/05/28)
この	•	イントロ		一般		配列取得		ゲノム配列		BSgenome (last modified 2014/06/28) NEW
すの	•	イントロ		一般		配列取得		プロモーター配列		公共DBから (last modified 2014/04/02)
201	•	イントロ		一般		配列取得		プロモーター配列		BSgenome (last modified 2014/04/25)
201	•	イントロ		一般		配列取得		プロモーター配列		GenomicFeatures(Lawrence 2013) (last modified 2014/06/25)
門	•	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 2014/04/02)
	•	イントロ		一般		配列取得		トランスクリプトーム配列		biomaRt(Durink 2009) (last modified 2014/06/25)
	•	イントロ		NGS		様々なプラットフォーム				(last modified 2014/06/10)
	•	イントロ		NGS		qPCRやmicroarrayなどとの比較				(last modified 2014/07/11) NEW
	•	イントロ		NGS		可視化(ゲノムブラウザやViewer)				(last modified 2014/06/25) NEW
	•	イントロ		NGS		配列取得		FASTQ or SRALite		公共DBから (last modified 2014/06/28) NEW
	•	イントロ		NGS		配列取得		FASTQ or SRALite		SRadb(Zhu 2013) (last modified 2014/06/25)
	•	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 2014/06/25) NEW
	•	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 2014/06/25)
	•	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 2014/04/11)
	•	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(Durink 2009)		(last modified 2013/09/26)
	•	イントロ		NGS		アノテーション情報取得		TranscriptDb		について (last modified 2014/03/28)
	•	イントロ		NGS		アノテーション情報取得		TranscriptDb		TxDb.*から (last modified 2013/10/08)
	•	イントロ		NGS		アノテーション情報取得		TranscriptDb		GenomicFeatures(Lawrence 2013) (last modified 2014/06/25)
	•	イントロ		NGS		アノテーション情報取得		TranscriptDb		GFF/GTF形式ファイルから (last modified 2014/06/25)

塩基配列データ取得



QCやpreprocessing



マッピング



カウント情報取得



発現変動解析

ヒトやマウスなどのリファレンス配列、NGSデータ、アノテーション情報取得などもR経由で可能。**wget**や**ftp**周辺



NGSデータ解析とR

(Rで)塩基配列解析

～NGS、RNA-seq、ゲノム、トランスクリプトーム、正規化、発現変動、統計、モデル、バイオインフォマティクス～
(last modified 2014/07/14, since 2010)

- イントロ | NGS | 読み込み | FASTA形式 | [基本情報を取得](#) (last modified 2014/05/29)
- イントロ | NGS | 読み込み | FASTA形式 | [description行の記述を整形](#) (last modified 2014/04/05)
- イントロ | NGS | 読み込み | FASTQ形式 | [FASTQ形式](#) (last modified 2014/06/15)
- イントロ | NGS | 読み込み | FASTQ形式 | [description行の記述を整形](#) (last modified 2013/06/13)
- イントロ | NGS | 読み込み | [Illuminaの * seq.txt](#) (last modified 2013/06/13)
- イントロ | NGS | 読み込み | [Illuminaの * qseq.txt](#) (last modified 2013/06/17)
- イントロ | [ファイル形式の変換](#) | [について](#) (last modified 2014/06/09)
- イントロ | ファイル形式の変換 | [BAM --> BED](#) (last modified 2014/06/21) **NEW**
- イントロ | ファイル形式の変換 | [FASTQ --> FASTA](#) (last modified 2013/06/17)
- イントロ | ファイル形式の変換 | [Genbank --> FASTA](#) (last modified 2014/03/10)
- イントロ | ファイル形式の変換 | [qseq --> FASTA](#) (last modified 2013/06/17)
- イントロ | ファイル形式の変換 | [qseq --> Illumina FASTQ](#) (last modified 2013/06/17)
- イントロ | ファイル形式の変換 | [qseq --> Sanger FASTQ](#) (last modified 2013/08/19)
- [前処理](#) | [クオリティチェック](#) | [について](#) (last modified 2014/06/30) **NEW**
- 前処理 | クオリティチェック | [qrgqc](#) (last modified 2014/06/11)
- 前処理 | クオリティチェック | [PHREDスコアに変換](#) (last modified 2013/06/18)
- 前処理 | クオリティチェック | [配列長分布を調べる](#) (last modified 2013/06/18)

塩基配列データ取得

QCやpreprocessing

マッピング

カウント情報取得

発現変動解析

FASTAやFASTQ形式ファイルの読み込み。ファイル形式の変換、Quality Control (QC)なども可能。**SAMtools**や**FastQC**周辺。



～NGS、RNA-seq、ゲノム、トランスクリプトーム、正規化、発現変動、統計、モデル、バイ
(last modified 2014/07/14, since 2010)

- FastQCなどR以外のプログラムもリストアップしています

R用:

- R以外:

- 19

NGSデータ解析とR

(Rで)塩基配列解析

～NGS、RNA-seq、ゲノム、トランスクリプトーム、正規化、発現変動、統計、モデル、バイオインフォマティクス～
(last modified 2014/07/14, since 2010)

- | What | この | すの | 201 | 201 | 門 | マッ | ニッ | した | 201 | 東 | 申 | 入 | 参 |
|------|---------|-----------------------------|----------------------------|----------------------------|---|----|----|----|-----|---|---|---|---|
| 前処理 | フィルタリング | PHREDスコアが低い塩基をNに置換 | (last modified 2014/03/03) | | | | | | | | | | |
| 前処理 | フィルタリング | PHREDスコアが低い配列(リード)を除去 | (last modified 2014/03/03) | | | | | | | | | | |
| 前処理 | フィルタリング | ACGTのみからなる配列を抽出 | (last modified 2013/06/18) | | | | | | | | | | |
| 前処理 | フィルタリング | ACGT以外の character "-" をNに変換 | (last modified 2013/06/18) | | | | | | | | | | |
| 前処理 | フィルタリング | ACGT以外の文字数が閾値以下の配列を抽出 | (last modified 2013/09/2) | | | | | | | | | | |
| 前処理 | フィルタリング | 重複のない配列セットを作成 | (last modified 2013/06/18) | | | | | | | | | | |
| 前処理 | フィルタリング | 指定した長さ以上の配列を抽出 | (last modified 2014/02/07) | | | | | | | | | | |
| 前処理 | フィルタリング | 指定した長さの範囲の配列を抽出 | (last modified 2013/06/18) | | | | | | | | | | |
| 前処理 | フィルタリング | 任意のIDを含む配列を抽出 | (last modified 2013/06/18) | | | | | | | | | | |
| 前処理 | フィルタリング | Illuminaの pass filtering | (last modified 2013/06/19) | | | | | | | | | | |
| 前処理 | フィルタリング | GFF/GTF形式ファイル | (last modified 2013/10/10) | | | | | | | | | | |
| 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配列 | (last modified 2014/06/11) | | | | | | | | | | |
| 前処理 | トリミング | ポリA配列除去 | ShortRead(Morgan 2009) | (last modified 2014/06/11) | | | | | | | | | |
| 前処理 | トリミング | アダプター配列除去(基礎) | girafe(Toedling 2010) | (last modified 2014/06/11) | | | | | | | | | |
| 前処理 | トリミング | アダプター配列除去(基礎) | ShortRead(Morgan 2009) | (last modified 2014/06/11) | | | | | | | | | |
| 前処理 | トリミング | アダプター配列除去(応用) | QuasR(Lerch 20XX) | (last modified 2014/06/13) | | | | | | | | | |
| 前処理 | トリミング | アダプター配列除去(応用) | ShortRead(Morgan 2009) | (last modified 2014/06/13) | | | | | | | | | |
| 前処理 | トリミング | 指定した末端塩基数だけ除去 | (last modified 2013/06/15) | | | | | | | | | | |

塩基配列データ取得



QCやpreprocessing



マッピング



カウント情報取得



発現変動解析

クオリティの低いリードの除去(フィルタリング)やアダプター配列の除去もできます。



NGSデータ解析とR

(Rで)塩基配列解析

～NGS、RNA-seq、ゲノム、トランスクリプトーム、正規化、発現変動、統計、モデル、バイオインフォマティクス～
(last modified 2014/07/14, since 2010)

- What
この
すの
201
ン
201
門
マッ
ニッ
した
201
東
申
ム
参
- [アセンブル](#) | [について](#) (last modified 2014/06/20) **NEW**
 - [アセンブル](#) | [ゲノム用](#) (last modified 2014/07/08) **NEW**
 - [アセンブル](#) | [トランスクリプトーム\(転写物\)用](#) (last modified 2014/07/08) **NEW**
 - [マッピング](#) | [について](#) (last modified 2014/07/09) **NEW**
 - [マッピング](#) | [basic aligner](#) (last modified 2014/07/09) **NEW**
 - [マッピング](#) | [splice-aware aligner](#) (last modified 2014/07/09) **NEW**
 - [マッピング](#) | [Bisulfite sequencing用](#) (last modified 2014/07/09) **NEW**
 - [マッピング](#) | [\(ESTレベルの長さの\)contig](#) (last modified 2014/06/24) **NEW**
 - [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\(Lerch XXX\)](#) (last modified 2013/06/19)
 - [マッピング](#) | [single-end](#) | [ゲノム](#) | [basic aligner\(応用\)](#) | [QuasR\(Lerch XXX\)](#) (last modified 2013/06/19)
 - [マッピング](#) | [single-end](#) | [ゲノム](#) | [splice-aware aligner](#) | [QuasR\(Lerch XXX\)](#) (last modified 2013/06/19)
 - [マップ後](#) | [について](#) (last modified 2013/06/19)
 - [マップ後](#) | [出力ファイル形式について](#) (last modified 2013/11/05)
 - [マップ後](#) | [出力ファイルの読み込み](#) | [BAM形式](#) (last modified 2014/06/21) **NEW**
 - [マップ後](#) | [出力ファイルの読み込み](#) | [Bowtie形式](#) (last modified 2013/06/18)
 - [マップ後](#) | [出力ファイルの読み込み](#) | [SOAP形式](#) (last modified 2013/06/19)
 - [マップ後](#) | [出力ファイルの読み込み](#) | [htSeqTools \(Planet 2012\)](#) (last modified 2013/06/19)
 - [マップ後](#) | [カウント情報取得](#) | [について](#) (last modified 2014/03/12)
 - [マップ後](#) | [カウント情報取得](#) | [ゲノム](#) | [アノテーション有](#) | [QuasR\(Lerch XXX\)](#) (last modified 2013/06/19)
 - [マップ後](#) | [カウント情報取得](#) | [ゲノム](#) | [アノテーション無](#) | [QuasR\(Lerch XXX\)](#) (last modified 2014/06/21)
 - [マップ後](#) | [カウント情報取得](#) | [トランスクリプトーム](#) | [BEDファイルから](#) (last modified 2014/06/21)
 - [マップ後](#) | [配列長とカウント数の関係](#) (last modified 2014/07/03) **NEW**

塩基配列データ取得



QCやpreprocessing



マッピング



カウント情報取得



発現変動解析

クオリティの低いリードの除去(フィルタリング)やアダプター配列の除去もできます。特にアダプター配列除去はsmall RNA-seqのマッピングに大きな影響を及ぼします。



- 解析 | 一般 | GC含量 (GC contents)(last modified 2014/05/01)
- 解析 | 一般 | Sequence logos(Schneider 1990) (last modified 2014/06/21) **NEW**
- 解析 | 一般 | 上流配列解析 | LDSS(Yamamoto 2007)(last modified 2012/07/17)
- 解析 | 一般 | 上流配列解析 | Relative Appearance Ratio(Yamamoto 2011)(last modified 2014/06/21)

• 解析 | 一般 | Sequence logos(Schneider 1990)

解析 | 一般 | Sequence logos(Schneider_1990) **NEW**

seqLogoパッケージを用いてsequence logos (Schneider and Stephens 1990)を実行する例を示します。ここでは

8. FASTQ形式ファイル(SRR609266.fastq.gz)の場合:

small RNA-seqデータ(400Mb弱、11,928,428リード)です。圧縮ファイルもreadDNAStringSet関数で通常手順で読み込みます。原著論文(Nie et al., BMC Genomics, 2013)中の記述から GSE41841を頼りに、SRP016842にたどりつき、[1. 入力 | NGS | 配列取得 | FASTQ or SRALite | SRADB\(Zhu 2013\)](#)の7を実行して得られたものが入力ファイルです。

1. 入力

```
in_f <- "SRR609266.fastq.gz"
out_f <- "hoge8.png"
param_fig <- c(800, 370)
```

#入力ファ
#出力ファ
#ファイル

small RNA-seqファイルをそのまま入力としてSequence logosを実行することもできる。つまりmultiple alignmentを行わないやり方

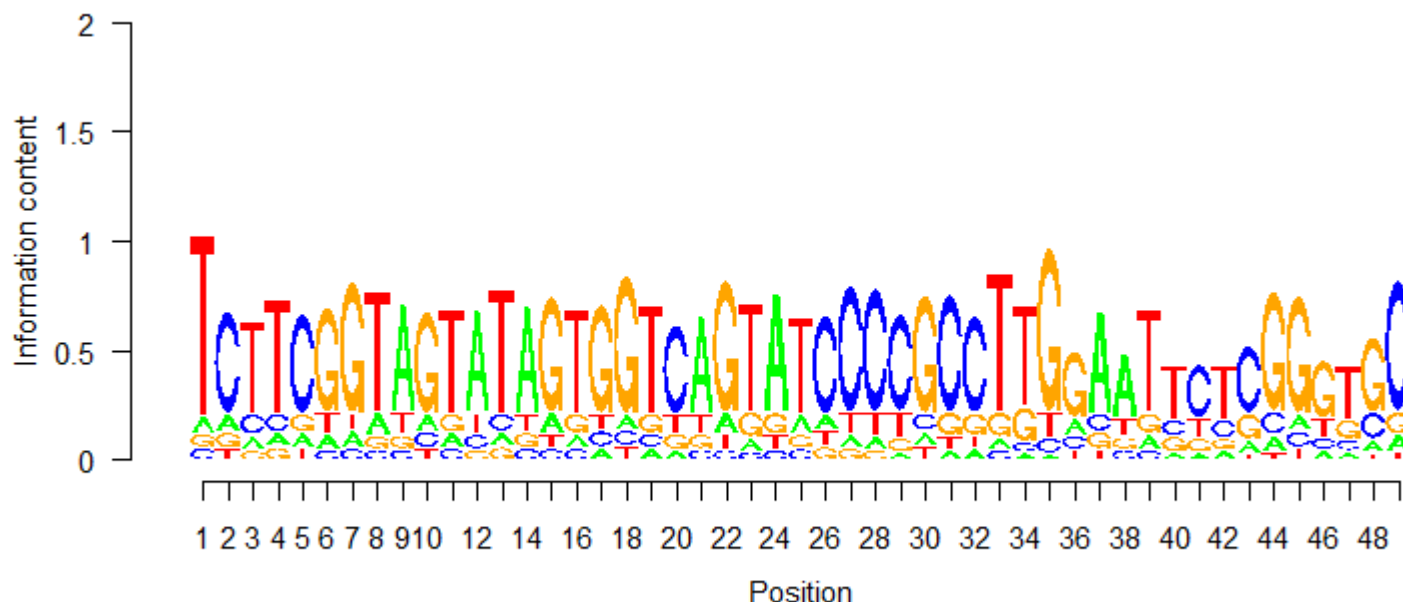
#必要なパッケージをロード

```
library(Biostrings)
library(seqLogo)
```

#入力ファ
fasta <-

#本番(seq
hoge <- c
out <- ma

#ファイル
png(out_f
seqLogo(c
dev.off())



解析 | 一般 | Sequence logos(Schneider_1990) **NEW**

seqLogoパッケージを用いてsequence logos (Schneider and Stephens 1990)を実行する例を示します。ここでは

8. FASTQ形式ファイル(SRR609266.fastq.gz)の場合:

small RNA-seqデータ(400Mb弱、11,928,428リード)です。圧縮ファイルもreadDNAStringSet関数で通常手順で読み込めます。原著論文(Nie et al., BMC Genomics, 2013)中の記述から [GSE41841](#)を頼りに、[SRP016842](#)にたどりつき、[1](#) [リンク | NGS | 配列取得 | FASTQ or SRALite | SRADB\(Zhu 2013\)](#)の7を実行して得られたものが入力ファイルです。

1. 入力

```

in_f <- "SRR609266.fastq.gz"
out_f <- "hoge8.png"
param_fig <- c(800, 370)
    
```

#入力ファ
 #出力ファ
 #ファイル

アダプター配列除去前のFASTQファイルでの実行結果。アダプター配列に相当する部分のロゴがよくわかる。

#必要なパッケージをロード

```

library(Biostrings)
library(seqLogo)
    
```

```

#入力ファ  

fasta <- readDNAStringSet(in_f)
    
```

```

#本番(seqLogo)を実行  

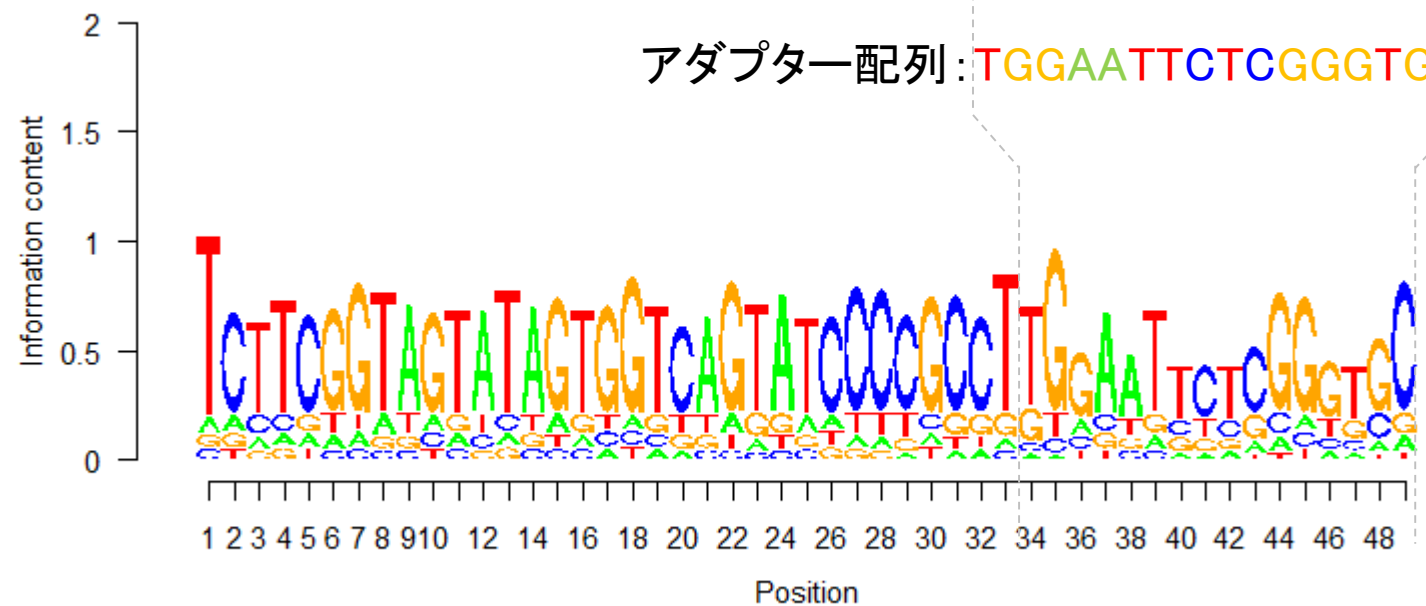
hoge <- seqLogo(fasta, param_fig)
out <- writePNG(hoge, out_f)
    
```

```

#ファイル名を指定して保存  

png(out_f, width=800, height=370, res=300)  

seqLogo(fasta, param_fig, dev.off())
    
```



解析 | 一般 | Sequence logos(Schneider_1990) NEW

seqLogoパッケージを用いてsequence logos (Schneider and Stephens 1990)を実行する例を示します。ここでは

9. FASTQ形式ファイル(hoge4.fastq.gz)の場合:

small RNA-seqデータ(280Mb弱、11,928,428リード)です。原著論文(Nie et al., BMC Genomics, 2013)中の記述からGSE41841を頼りに、SRP016842にたどりつき、前処理 | トリミング | アダプター配列除去(応用) | ShortRead (Morgan 2009)の4を実行して得られたものが入力ファイルです。アダプター配列除去後のデータなので、リードごとに配列長が異なる場合でも読み込めるShortReadパッケージ中の

アダプター配列除去後のFASTQファイルでの実行結果。アダプター配列に相当する部分のロゴが消えていることがわかる。

```

in_f <- "hoge4.fastq.gz"
out_f <- "hoge9.png"
param_fig <- c(787, 370)

```

#入力ファ
 #出力ファ
 #ファイル

```

#必要なパ
library(S
library(s

```

```

#入力ファ
fastq <-
fasta <-

```

```

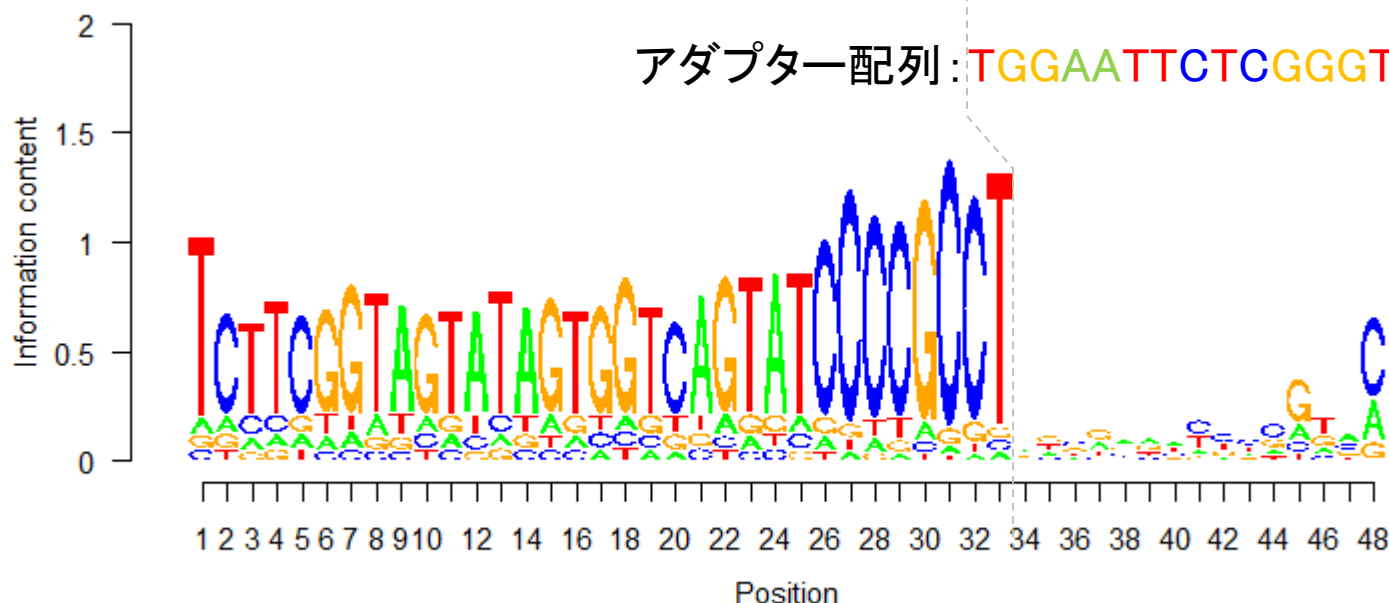
#本番(seq
hoge <- c
out <- ma

```

```

#ファイル
png(out_f
seqLogo(c
dev.off()

```



NGSデータ解析とR

■ 発現変動解析

□ 入力: カウントデータ

- 遺伝子発現行列のような数値行列
- 整数値からなる遺伝子領域上にマップされたリード数

□ 出力: 発現変動遺伝子リスト (p -value や q -value) や M-A plot

塩基配列データ取得



QCやpreprocessing



マッピング



カウント情報取得



発現変動解析

入力: カウントデータ

	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG000000000003	480	513	366	124	271	366
ENSG000000000005	0	0	0	1	0	0
...						
ENSG00000240386	0	0	0	4001	5500	6851
...						
ENSG00000128564	18	27	19	2038	2657	2138
...						

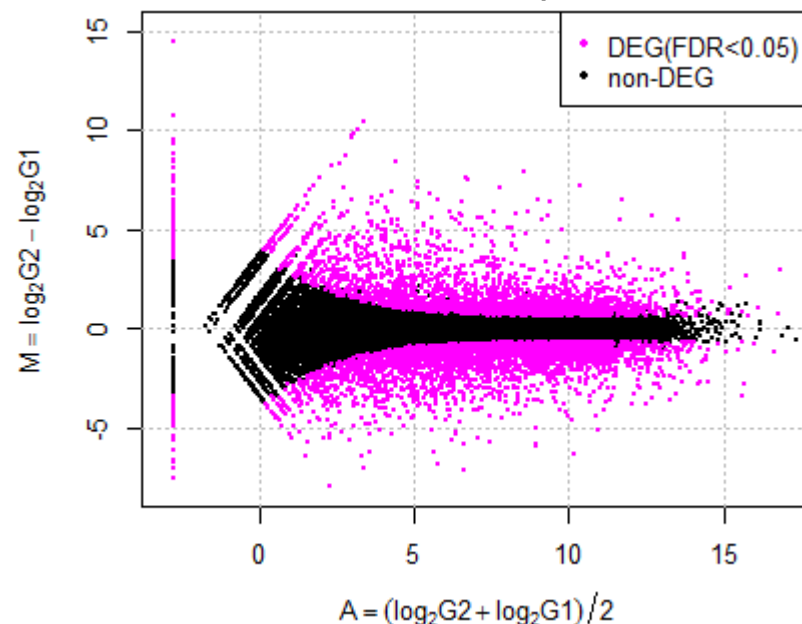
G1群

G2群

発現変動
解析



出力: M-A plot



マッピング → カウントデータ取得 → 発現変動解析も可能

謝辞

まとめ: Rでもいろいろできます

共同研究者

清水 謙多郎 先生(東京大学・大学院農学生命科学研究科)

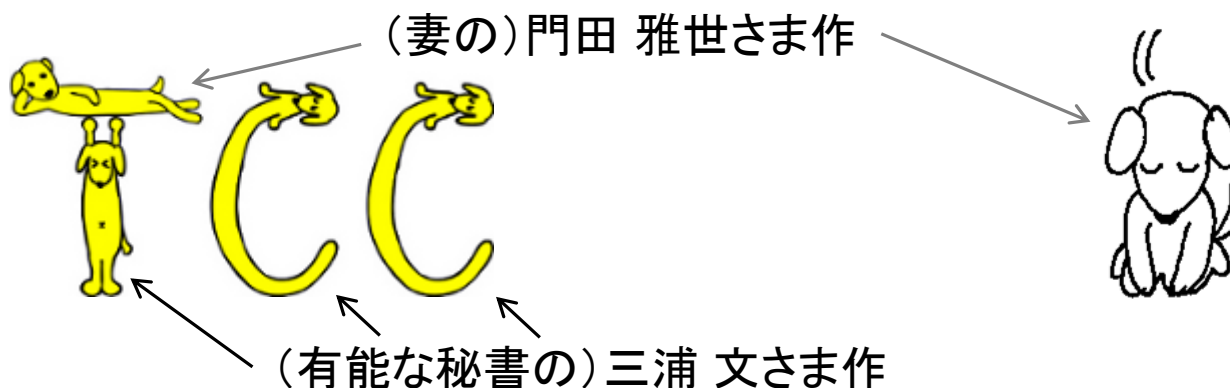
西山 智明 先生(金沢大学・学際科学実験センター)

孫 建強 氏(東京大学・大学院農学生命科学研究科・大学院生)

グラント

- 基盤研究(C)(H24-26年度):「シーケンスに基づく比較トランスクリプトーム解析のためのガイドライン構築」(代表)
- 新学術領域研究(研究領域提案型)(H22年度-):「非モデル生物におけるゲノム解析法の確立」(分担; 研究代表者: 西山智明)

挿絵やTCCのロゴなど



NGSデータ解析とR

塩基配列データ取得

QCやpreprocessing

マッピング

カウント情報取得

発現変動解析

- 解析 | 発現変動 | について (last modified 2014/07/10) **NEW**
- 解析 | 発現変動 | 2群間 | 対応なし | について (last modified 2014/07/10) **NEW**
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC (Sun 2013) (last modified 2014/07/10)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | edgeR (Robinson 2010) (last modified 2014/07/10)

解析 | 発現変動 | 2群間 | 対応なし | について **NEW**

実験デザインが以下のような場合にこのカテゴリーに属する方法を適用します:

Aさんの正常サンプル
Bさんの正常サンプル
Cさんの正常サンプル
Dさんの腫瘍サンプル
Eさんの腫瘍サンプル
Fさんの腫瘍サンプル
Gさんの腫瘍サンプル

2014年7月に調査した結果

R用:

- DEGSeq: Wang et al., Bioinformatics, 2010
- edgeR: Robinson et al., Bioinformatics, 2010
- GPseq: Srivastava et al., Nucleic Acids Res., 2010
- baySeq: Hardcastle and Kelly, BMC Bioinformatics, 2010
- DESeq: Anders and Huber, Genome Biol., 2010

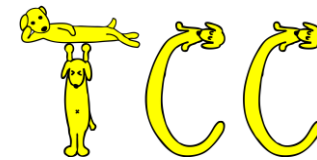
R用:

- DEGSeq: Wang et al., Bioinformatics, 2010
- edgeR: Robinson et al., Bioinformatics, 2010
- GPseq: Srivastava et al., Nucleic Acids Res., 2010
- baySeq: Hardcastle and Kelly, BMC Bioinformatics, 2010
- DESeq: Anders and Huber, Genome Biol., 2010
- DESeq2: Anders and Huber, Genome Biol., 2010
- NBPSeq: Di et al., SAGMB, 2011
- BBSeq: Zhou et al., Bioinformatics, 2011
- NOISeq: Tarazona et al., Genome Res., 2011
- PoissonSeq: Li et al., Biostatistics, 2012
- SAMseq: Li and Tibshirani, Stat Methods Med Res., 2012
- easyRNASeq: Delhomme et al., Bioinformatics, 2012
- DSGseq: Wang et al., Gene, 2013
- sSeq: Yu et al., Bioinformatics, 2013
- TCC: Sun et al., BMC Bioinformatics, 2013
- tweeDEseq: Esnaola et al., BMC Bioinformatics, 2013
- NPEBseq: Bi et al., BMC Bioinformatics, 2013
- DER Finder: Frazee et al., Biostatistics, 2014
- Characteristic Direction(CD): Clark et al., BMC Bioinformatics, 2014
- edgeR-robust: Zhou et al., Nucleic Acids Res., 2014
- ShrinkBayes: Van De Wiel et al., BMC Bioinformatics, 2014

最も有名なのはedgeRとDESeq

我々はTCCを提供

発現変動解析用パッケージ



■ TCC (Sun et al., *BMC Bioinformatics*, 2013)

- TCCは内部的に既存パッケージ(edgeR, DESeq, and baySeq)中の関数を利用。既存パッケージ中のオリジナルの手順を繰り返し実行することで、データ正規化精度向上を実現。オリジナルの手順のみの場合に比べてより感度・特異度の高いDEG検出結果を得ることができる。
- TCC原著論文中では、edgeR, DESeq, baySeqパッケージ中の関数を自在に組み合わせて実行し、2群間比較の場合のみで性能評価している。推奨は以下の通り:
 - Biological replicatesありの場合: edgeR中の関数のみからなるiDEGES/edgeR正規化法
 - Biological replicatesなしの場合: DESeq中の関数のみからなるiDEGES/DESeq正規化法
- 実質的には、より頑健なiterative edgeRやiterative DESeqを簡単に実行できるパッケージがTCCという理解で差支えない。
- 2013年7月の論文publish以降も継続的にアップデートしています
 - 多群間比較やpaired dataへの対応など、解析可能な実験デザインを拡張
 - DESeq2対応もほぼ完了
 - サンプル間クラスタリング用関数やマイクロアレイデータ用組織特異的発現パターン検出法ROKUの実装
 - ドキュメントが充実(TCC ver. 1.4.0で74ページに!)



compcodeRによる性能評価
でもTCCの優位性を確認済