



# バイオインフォマティクス ～LinuxでNGS解析(の基礎)～

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](mailto:kadota@iu.a.u-tokyo.ac.jp)

<http://www.iu.a.u-tokyo.ac.jp/~kadota/>

# Contents

## ■ イン트로ダクション

- 概要、背景(NGS用カリキュラム、講習会)、Linuxスキル習得の意義
- ウェブ情報(日本乳酸菌学会誌のNGS連載やNGS講習会資料)

## ■ 実習環境に慣れる

- 仮想環境での作業に慣れる
- GUIとCUI(マウス操作かコマンド入力操作か)
- ターミナルでの作業
- 共有フォルダの概念を理解

## ■ 練習

- 作業ディレクトリの変更、練習用NGSデータファイルのダウンロード
- ファイルの確認、de novoゲノムアセンブリ
- BLAST検索

## ■ 課題

- グループごとに異なる課題ファイルを入力として、「ダウンロード、de novoアセンブリ、BLAST検索」を実行し、得られた結果をレポートにまとめて発表せよ
- グループ1はkadai1.fasta、グループ2はkadai2.fasta、etc.

(主にNGS解析を意識した)バイオインフォマテイクススキルの習得がメインだが、何かをやったという達成感も得られるように実際のNGSデータの一部を用いてゲノムアセンブリまで行う

# 概要

## ■ キーワード

- NGS, Linux, バイオインフォマテイクス, 仮想環境, Bio-Linux, ゲノムアセンブリ

## ■ Linux

- WindowsやMacintoshと同じく、OSの一種
- バイオインフォマテイクス分野でよく利用される
- 「Windowsのコマンドプロンプト」や「Macintoshのターミナル」と同じく、lsやcdなどのLinuxコマンドを知らなければ何もできないため、慣れるまでが大変
- 使いこなせれば、最先端の解析用プログラムを自在にインストール・利用可能となり、効率的かつ**通り一辺倒でない**データ解析も可能となる

## ■ 次世代シーケンサ(NGS)解析

- NGSとは、大量の塩基配列を出力する実験機器(Next-Generation Sequencer)またはその技術を指す。主にゲノム解析やトランスクリプトーム解析と呼ばれる分野で利用されている
- 塩基配列解析用プログラムは、UNIX(今のLinux)環境で動作するものが多かった歴史的背景などから、現在でもLinux上で動くプログラムがまず最初に開発される場合が多い

# 背景

「R NGS」などでググリ、①のウェブページへ。②または③のあたりをクリック

## (Rで)塩基配列解析

(last modified 2016/08/24, since 2011)

このウェブページは [インストール](#) についての推奨手順 ([Windows2015.04.04版](#)と[Macintosh2015.04.03版](#))に従ってフリーソフトRと必要なパッケージをインストール済みであるという前提で記述しています。初心者の方は [基本的な利用法](#) ([Windows2015.04.03版](#)と[Macintosh2015.04.03版](#))で自習してください。本ウェブページを体系的にまとめた [書籍](#) もあります。(2015/04/03)

### What's new?

- バイオインフォマティクスやNGSをキーワード程度は知っているヒト(学部3年生程度)向けの講義資料を作成しました。「参考資料 | [講習会](#)、[講義](#)、[講演資料](#)」の2016.09.12-16の日付のものです。[日本乳酸菌学会誌のNGS連載](#)第1-4回あたりのダイジェスト版のようなものです。[NGSハンズオン講習会2016](#)の予習事項が厳しすぎて受講を断念したヒトも、ここからだとスムーズに頭に入っていかも知れません。(2016/08/24) **NEW**
- 参考資料の項目の情報量が多くなってきたので、「[書籍](#)、[学会誌](#)」と「[講習会](#)、[講義](#)、[講演資料](#)」の2つに分割しました。(2016/08/23) **NEW**
- [NGSハンズオン講習会2016](#)の「昨年度との違い、対応関係、想定受講者、および予習事項」の [PDF](#) に注釈を入れました。8/4の cuffdiff の入力ファイル周辺の話です。経緯などの記憶は定かではありませんが、ここで誤解を与えていたかもしれないと思いましたので、念のため追記しました。これ以外にも間違いや私の誤解などありましたら、よりよい情報共有のために遠慮なくご指摘よろしくお願いたします m(\_ \_)m(2016/08/19) **NEW**
- [NGSハンズオン講習会2016](#)の第3部(8/4)の [講義資料PDF](#) を修正しました。スライド107 (Trinityのパスを通すところのミス)、スライド140 (Bridgerに必要なboostパッケージをapt-getでやるとサンプルデータの実行もうまくいく)あたりです。(2016/08/12) **NEW**
- [NGSハンズオン講習会2016](#)の第2部(7/25-28)の講義資料修正版を公開しました。(2016/08/03) **NEW**

- [門田からメール返信をもらえない場合は](#) (last modified 2016/08/23) **NEW**
- [はじめに](#) (last modified 2015/03/31)
- [参考資料 | 書籍、学会誌](#) (last modified 2016/08/23) **NEW**
- [参考資料 | 講習会、講義、講演資料](#) (last modified 2016/08/24) **NEW**
- [過去のお知らせ](#) (last modified 2016/08/24) **NEW**

[トップページへ](#)

# 背景

①2016.09-12-16の講義資料に辿りつく。ここは、私の講習会、講義、講演資料が公開されています。②をクリック

## 参考資料 | 講習会、講義、講演資料 NEW

基本的に私門田の個人ページに記載してあるものです。かなり古い講演資料などの情報をもとに勉強されている方もいらっしゃるようですので、ここでは2013年秋以降の情報のみです。大まかな内容についても述べています。講義資料としての利用などは事前連絡や私個人への謝辞も気にせずご自由にお使いください。

- 門田幸二「[講義資料](#)」, 東京大学農学部展開科目: バイオインフォマティクス, 東京大学(東京), 2016.09.12-16

内容: バイオインフォマティクスやNGSをキーワード程度は知っている(学部3年生程度)向けのアクティブ・ラーニング系講義。バイオインフォマティクスとLinuxスキル習得の意義。バイオインフォマティクス分野で需要が多いのはNGSデータの解析 (JST-NBDCによる[アンケート結果](#))。バイオインフォマティクス人材育成のための講習会の一環として行われているNGS講習会資料を有効活用するために必要な基本スキルの習得。具体的には、日本乳酸菌学会誌のNGS連載内容を理解するための基礎として、[Bio-Linux](#)のノリに慣れるのが最低限の到達目標。何かやったという達成感を味わってもらうため、NGSデータのde novoゲノムアセンブリを行い、元のリード長(塩基配列の長さ)よりも伸びたことを実感してもらう。また、アセンブリ結果の一部をBLAST検索し、入力データ([hoge.fasta](#); 約6.4MB)がどんな生物種であったかを調べるあたりまで。課題は[kadai1.fasta](#), [kadai2.fasta](#), [kadai3.fasta](#), [kadai4.fasta](#)のいずれかを選択して実行。下記と同様の手順で「データのダウンロード → de novoアセンブリ → BLAST検索」し、得られた結果を発表。約2日分。

#作業ディレクトリの変更

```
cd /home/iu/Desktop/mac_share
```

#解析したいファイル([hoge.fasta](#))のダウンロード

```
wget -c http://www.iu.a.u-tokyo.ac.jp/%7Ekadota/20160912/hoge.fasta
```

#行数やファイルサイズを表示(行数が200,000行、サイズが6,688,895 bytesとなっていればOK)

```
wc hoge.fasta
```

#最初の10行分を表示(2行分で1つのリードを表し、1リードが50塩基であることを確認)

```
head hoge.fasta
```

#de novoアセンブリを実行し、結果をugeディレクトリに保存

```
velveth uge 31 -short -fasta hoge.fasta
```

```
velvetg uge
```

#ugeディレクトリに移動して、アセンブリ結果ファイル(contigs.fa)があることを確認

```
cd uge
```

[トップページへ](#)

# 背景

①需要の多い次世代シーケンサ(Next-Generation Sequencer; NGS)から得られる大量塩基配列データを効率的に解析するためのバイオインフォマティクス人材育成カリキュラム(NGS用カリキュラム)。平成26年3月公開

## バイオインフォマティクス人材育成のための講習会

### NGS解析

ライフサイエンス分野の研究現場においては、莫大・多様な研究データが産出され、取り扱うデータ量が飛躍的に増えています。一方で、それらのデータを整備・活用するための人材は不足している状況です。そのような研究現場の状況を踏まえ、NBDC運営委員会人材育成分科会において、研究データを整備・活用するバイオインフォマティクス人材を育成するためのカリキュラムに基づき実施した講習会です。本カリキュラムは、対応が急務であると思われる次世代シーケンサデータに焦点をあてた内容となっております。

#### ●人材育成分科会で策定したカリキュラム

[バイオインフォマティクス人材育成カリキュラム \(次世代シーケンサ\)](#)

①

[カリキュラムで習得できる技能](#)

[カリキュラム フロー図](#)

#### ●H28年度NGSハンズオン講習会 (2016年7月19日～8月4日)

※受講者アンケートにご協力をお願いいたします (予めメールでお知らせした認証IDとパスワードをご用意ください)。

[受講者アンケート回答はこちら](#)

回答期間: 7/19 (火) ~9/16 (金)

#### ●H27年度NGSハンズオン講習会 (2015年7月22日～8月6日)

#### ●H26年度NGS速習コース講習会 (2014年9月1日～12日)

# NGS用カリキュラム

NGS用カリキュラムの中身。NGSデータ解析に最低限必要とされる知識・技術を2週間程度で身につけることを想定した「速習」と、時間をかけて習得することを想定した「速習以外」にわかれている。ここで示しているのは①「速習」

【速習】

大項目	日数	No.	項目	習得技術		
1. コンピュータリテラシーとサーバー設計	4日	2日	1-1 OS, ハード構成	・コンピュータの基本の理解		
			1-2 ネットワーク基礎	・インターネット、セキュリティの基本の理解	初級	講義
			1-3 UNIX I	・UNIXの基礎の理解 ・Linux導入	中級	実習
	2日	1-4 スクリプト言語	・Perl ・シェルスクリプト	中級	実習	
2. 配列インフォマティクス	1日	2-1 配列解析基礎	・配列、ゲノムデータ記述のフォーマット、アラインメント(DP)、データベース検索(BLAST, BLAT)等の基礎的な配列比較解析の原理と実習	初級	実習	
		2-2 バイオ系データベース概論	・基本的な各種バイオ系データベースの理解、統合DBの利用法	初級	実習	
3. データ解析基礎	2日	3-1 R 基礎1	・R言語の基礎(インストールから利用まで)	初級	実習	
		3-2 R 基礎2	・ファイルの読み込み、行列演算の基本	初級	実習	
		3-3 R 各種パッケージ	・Rの各種パッケージのインストール法と代表的なパッケージの利用法	中級	実習	
		3-4 R bioconductor I	・bioconductorの利用法	中級	実習	
		3-5 R bioconductor II	・FASTA and FASTQ形式ファイルの読み込み。ファイル形式の変換(FASTQ → FASTA)、クオリティチェック、リード配列長分布、フィルタリングやトリミング、GC含量計算など	中級	実習	
4. 次世代シーケンサ	0.25日	4-1 次世代シーケンサ基礎I	・原理の理解	初級	講義	
		4-2 次世代シーケンサ基礎II	・応用分野とそのための計測技術の理解(RNA-seq, ChIP-seq, がんゲノム, 個人ゲノム, 環境ゲノム, Hi-C)	初級	講義	
	0.5日	4-3 次世代シーケンサ実習I	・ファイル形式、可視化、quality check、マッピング、アセンブル	初級	実習	
	1.5日	4-4 次世代シーケンサ実習II	・代表的なパイプラインについての実習(多型解析(IGV)、RNA-seq、ChIP-seq、及び統合解析、定番のツールを利用)	初級	実習	
5. ゲノム関連の倫理・法律	0.5日	5-1	ゲノム情報倫理概論	・ゲノム情報を扱う上で、プライバシー保護などの必要な倫理的問題、法的問題の国内外の状況を理解し、ゲノム情報を適切に利用できるようにする ・匿名化、暗号化、情報セキュリティ概要	初級	講義
6. 分子生命科学	0.5日	6-1	分子生命科学概論	・複製、転写、翻訳、代謝、シグナル伝達などの基礎知識	初級	講義
		6-2	オミクス概論	・ゲノム以外のオミクスデータの基礎知識	初級	講義
		6-3	遺伝/進化概論	・ゲノムデータを扱う上での遺伝学、進化学の基礎知識	初級	講義

# NGS用カリキュラム

【速習以外】

大項目	No.	項目	習得技術	レベル	形式
1. コンピュータリテラシーとサーバー設計	1-5	UNIX II	・Windows OS上のUNIX環境構築	中級	実習
	1-6	プログラミング	・C, C++, Java	中級	講義
	1-7	ウェブ開発技術 I	・サーバ構築(apache, XML, JavaScript, Ajaxなど)、公開に関する技術	中級	実習
	1-8	並列計算技術	・並列計算プログラミング技術大規模データ解析のために利用することができるクラスターマシン、クラウド環境(遺伝研スパコン、amazon EC2など)の選定 ・遺伝研スパコンなどの大型計算機システム利用のためのキューイングシステム活用法	上級	講義
	1-9	ウェブ開発技術 II	・ビッグデータに対応したサーバ構築技術(Hadoopなど)	上級	実習
2. 配列インフォマティクス	2-3	分子系統解析	・分子系統解析ツールの導入や解析技術	中級	実習
	2-4	遺伝子機能解析	・COG,KEGG,GOなどを利用して、遺伝子配列から遺伝子機能の推定方法	中級	実習
4. 次世代シーケンサ	4-5	次世代シーケンサ 最新技術	・企業からの最新情報 ・機器のスペック、適した利用例及び標準的な導入価格の紹介	中級	講義
	4-6	アセンブル、マッピング	・次世代シーケンサデータ処理技術 ・アセンブル、マッピング	中級	実習
	4-7	発現解析	・次世代シーケンサデータ処理技術 ・RNA-seq 発現解析 ・BAMファイルの読み込み ・BAM → BED形式への変換 ・カウント情報取得(リファレンス配列の違い、アノテーション情報利用の有無など)	中級	実習
	4-8	クロマチン構造解析	・次世代シーケンサデータ処理技術 ・ChIP-seq解析技術 ・Hi-C	中級	実習
	4-9	環境ゲノム解析	・次世代シーケンサデータ処理技術 ・環境ゲノム解析技術の導入および利用技術 ・環境ゲノム解析技術 ・16S rRNAデータ解析: RDPデータベースやqiimeを利用した系統組成推定 ・メタゲノム解析: COG,KEGGを利用した機能組成推定	中級	実習
	4-10	比較ゲノム解析	・次世代シーケンサデータ処理技術 ・比較ゲノム解析 ・ゲノム内比較、種間比較に必要なツール(BLAST, HMMER, MAFFTなどを含む)の導入や解析技術	上級	実習
	4-11	個人ゲノム解析	・次世代シーケンサデータ処理技術 ・個人ゲノム解析 ・正の選択検出、SNVのリスク予測のツールの導入や解析技術	上級	実習
	4-12	疾患ゲノム解析	・次世代シーケンサデータ処理技術 ・CNV解析 ・エキソーム解析 ・変異同定	上級	実習
5. ゲノム関連の倫理・法律	5-2	ゲノム情報倫理応用	・匿名化、暗号化、情報セキュリティ詳細(技術的内容など)	上級	講義

# NGS用カリキュラム

NGS用カリキュラムの中身。NGSデータ解析に最低限必要とされる知識・技術を2週間程度で身につけることを想定した①「速習」の内容をとりあえずやってみたのが...

【速習】

大項目	日数	No.	項目	習得技術		
1. コンピュータリテラシーとサーバー設計	4日	2日	1-1 OS, ハード構成	・コンピュータの基本の理解	初級	講義
			1-2 ネットワーク基礎	・インターネット、セキュリティの基本の理解	初級	講義
			1-3 UNIX I	・UNIXの基礎の理解 ・Linux導入	中級	実習
	2日	1-4 スクリプト言語	・Perl ・シェルスクリプト	中級	実習	
2. 配列インフォマティクス	1日	2-1 配列解析基礎	・配列、ゲノムデータ記述のフォーマット、アラインメント(DP)、データベース検索(BLAST, BLAT)等の基礎的な配列比較解析の原理と実習	初級	実習	
		2-2 バイオ系データベース概論	・基本的な各種バイオ系データベースの理解、統合DBの利用法	初級	実習	
3. データ解析基礎	2日	3-1 R 基礎1	・R言語の基礎(インストールから利用まで)	初級	実習	
		3-2 R 基礎2	・ファイルの読み込み、行列演算の基本	初級	実習	
		3-3 R 各種パッケージ	・Rの各種パッケージのインストール法と代表的なパッケージの利用法	中級	実習	
		3-4 R bioconductor I	・bioconductorの利用法	中級	実習	
		3-5 R bioconductor II	・FASTA and FASTQ形式ファイルの読み込み。ファイル形式の変換(FASTQ → FASTA)、クオリティチェック、リード配列長分布、フィルタリングやトリミング、GC含量計算など	中級	実習	
4. 次世代シーケンサ	0.25日	4-1 次世代シーケンサ基礎I	・原理の理解	初級	講義	
		4-2 次世代シーケンサ基礎II	・応用分野とそのための計測技術の理解(RNA-seq, ChIP-seq, がんゲノム, 個人ゲノム, 環境ゲノム, Hi-C)	初級	講義	
	0.5日	4-3 次世代シーケンサ実習I	・ファイル形式、可視化、quality check、マッピング、アSEMBル	初級	実習	
	1.5日	4-4 次世代シーケンサ実習II	・代表的なパイプラインについての実習(多型解析(IGV)、RNA-seq、ChIP-seq、及び統合解析、定番のツールを利用)	初級	実習	
5. ゲノム関連の倫理・法律	0.5日	5-1	ゲノム情報倫理概論	・ゲノム情報を扱う上で、プライバシー保護などの必要な倫理的問題、法的問題の国内外の状況を理解し、ゲノム情報を適切に利用できるようにする ・匿名化、暗号化、情報セキュリティ概要	初級	講義
6. 分子生命科学	0.5日	6-1	分子生命科学概論	・複製、転写、翻訳、代謝、シグナル伝達などの基礎知識	初級	講義
		6-2	オミクス概論	・ゲノム以外のオミクスデータの基礎知識	初級	講義
		6-3	遺伝/進化概論	・ゲノムデータを扱う上での遺伝学、進化学の基礎知識	初級	講義

# NGS速習コース講習会

## バイオインフォマティクス人材育成のための講習会

### NGS解析

ライフサイエンス分野の研究現場においては、莫大・多様な研究データが産出され、取り扱うデータ量が飛躍的に増えています。一方で、それらのデータを整備・活用するための人材は不足している状況です。そのような研究現場の状況を踏まえ、NBDC運営委員会人材育成分科会において、研究データを整備・活用するバイオインフォマティクス人材を育成するためのカリキュラムに基づき実施した講習会です。本カリキュラムは、対応が急務であると思われる次世代シーケンサデータに焦点をあてた内容となっております。

#### ●人材育成分科会で策定したカリキュラム

[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）](#)

[カリキュラムで習得できる技能](#)

[カリキュラム フロー図](#)

#### ●H28年度NGSハンズオン講習会（2016年7月19日～8月4日）

※受講者アンケートにご協力をお願いいたします（予めメールでお知らせした認証IDとパスワードをご用意ください）。

[受講者アンケート回答はこちら](#)

回答期間：7/19（火）～9/16（金）

#### ●H27年度NGSハンズオン講習会（2015年7月22日～8月6日）

#### ●H26年度NGS速習コース講習会（2014年9月1日～12日）

①

# NGS速習コース講習会

①カリキュラム通りに行ったので、座学(講義)のみの時間もあつた。また、計10日間にもおよぶため②担当講師数も多く連携をとりきれなかった。結果として③報告書中の受講生アンケートの主な要望は「**実習**のみで全体の連携」をとってほしい、であつた

■平成26年度NGS速習コース講習会(2014年9月1日~12日)

「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)」の速習部分について、講習会を開催しました。

H26年度概要

H26年度講義日程・参考資料

H26年度講習会時の情報については[こちら](#)もご覧ください。

H26年度実施報告書・講義資料・動画等

●講習会実施報告書(PDFファイル:3.7MB)

③

●講義資料・動画

①

②

実施日	実施時間	大項目	項目番号および項目	レベル	形式	担当講師(敬称略)	講義資料・動画
9月1日	10:40-12:00	1.コンピュータリテラシーとサーバー設計	1-1. OS、ハード構成	初級	講義	中村 保一 (DDBJ)	統合TV
	13:15-14:45		1-2. ネットワーク基礎				
	15:00-16:30		1-3. UNIX I	初級	実習	仲里 猛留 (DBCLS)	統合TV
	16:45-18:15						
9月2日	10:30-12:00			中級	実習	仲里 猛留 (DBCLS)	統合TV
	13:15-14:45						
	15:00-16:30						
	16:45-18:15						
9月3日	10:30-12:00		1-4. スクリプト言語	中級	実習	服部 恵美(アメリカエ)	統合TV
	13:15-14:45						

# NGSハンズオン講習会

①平成27年7-8月に行われた「NGSハンズオン講習会」では、実習に特化した内容で実施

## バイオインフォマティクス人材育成のための講習会

### NGS解析

ライフサイエンス分野の研究現場においては、莫大・多様な研究データが産出され、取り扱うデータ量が飛躍的に増えています。一方で、それらのデータを整備・活用するための人材は不足している状況です。そのような研究現場の状況を踏まえ、NBDC運営委員会人材育成分科会において、研究データを整備・活用するバイオインフォマティクス人材を育成するためのカリキュラムに基づき実施した講習会です。本カリキュラムは、対応が急務であると思われる次世代シーケンサデータに焦点をあてた内容となっております。

#### ●人材育成分科会で策定したカリキュラム

[バイオインフォマティクス人材育成カリキュラム \(次世代シーケンサ\)](#)

[カリキュラムで習得できる技能](#)

[カリキュラム フロー図](#)

#### ●H28年度NGSハンズオン講習会 (2016年7月19日～8月4日)

※受講者アンケートにご協力をお願いいたします (予めメールでお知らせした認証IDとパスワードをご用意ください)。

[受講者アンケート回答はこちら](#)

回答期間 : 7/19 (火) ~9/16 (金)

#### ●H27年度NGSハンズオン講習会 (2015年7月22日～8月6日)

①

#### ●H26年度NGS速習コース講習会 (2014年9月1日～12日)

# NGSハンズオン講習会

①Linux基礎の項目は1日分しかないが、1日でLinuxの基礎を習得可能というわけではない！

## 平成27年度NGSハンズオン講習会

平成27年度は、平成26年度の実績を踏まえ、講義内容の改善等を行い、ハンズオンに特化した、より効果的なNGS講習会を開催しました。

### H27年度概要

### H27年度講義日程・参考資料

H26年度講習会の情報については[こちら](#)をご覧ください。

### H27年度実施報告書・講義資料・動画等

● [講習会実施報告書 \(PDF: 2.17MB\)](#) および [受講者アンケート集計結果 \(データ集\) \(PDF: 662KB\)](#)

● [講義資料・動画](#) \*講義資料一覧のファイル名をクリックすると資料ファイル (PDF等) がダウンロードできます。

実施日	実施時間	大項目	項目	レベル	習得技術	担当講師(敬称略)	講義資料・動画(統合TV)
7月22日 (水)	10:30-12:00	PC環境の構築	Bio-Linux8とRのインストール状況確認		・Linux導入 ・R導入 ・NGS解析に必要な環境構築技術	門田 幸二 (東京大学)  寺田 透 (東京大学)	<a href="#">事前予習資料一覧</a> (PDF:52KB)
	13:15-14:45						
	15:00-16:30						<a href="#">講義資料一覧</a> (PDF:108KB)
	16:45-18:15						
7月23日 (木)	10:30-12:00	UNIX/Linuxとスクリプト言語	Linux基礎	初級	UNIXの基礎の理解	門田 幸二 (東京大学)	<a href="#">講義資料一覧</a> (PDF:32KB)
	13:15-14:45			中級			
	15:00-16:30			<a href="#">統合TV</a>			
	16:45-18:15						
7月24日 (金)	10:30-12:00		スクリプト言語	中級	シェルスクリプト	服部 恵美 (アメリエフ)	<a href="#">講義資料</a> (PDF:1.8MB)
	13:15-14:45						



# NGSハンズオン講習会

①Linux基礎は、②事前予習事項の復習という位置づけ。講習会受講者の大半は、(Windows上で)Linuxコマンドを利用可能な③Bio-Linux8という解析環境を自力で構築するところからスタートして、1週間程度はかかる自習をしてきたヒト

## 平成27年度NGSハンズオン講習会

平成27年度は、平成26年度の実績を踏まえ、講義内容の改善等を行い、ハンズオンに特化した、より効果的なNGS講習会を開催しました。

### H27年度概要

### H27年度講義日程・参考資料

H26年度講習会の情報については[こちら](#)をご覧ください。

### H27年度実施報告書・講義資料・動画等

● [講習会実施報告書 \(PDF: 2.17MB\)](#) および [受講者アンケート集計結果 \(データ集\) \(PDF: 662KB\)](#)

● [講義資料・動画](#) \*講義資料一覧のファイル名をクリックすると資料ファイル (PDF等) がダウンロードできます。

実施日	実施時間	大項目	項目	レベル	習得技術	担当講師(敬称略)	講義資料・動画(統合TV)	
7月22日 (水)	10:30-12:00	PC環境の構築	Bio-Linux8とRのインストール状況確認		<ul style="list-style-type: none"> <li>Linux導入</li> <li>R導入</li> <li>NGS解析に必要な環境構築技術</li> </ul>	門田 幸二 (東京大学)	<a href="#">事前予習資料一覧 (PDF:52KB)</a>	
	13:15-14:45						寺田 透 (東京大学)	<a href="#">講義資料一覧 (PDF:108KB)</a>
	15:00-16:30							
	16:45-18:15							
7月23日 (木)	10:30-12:00	UNIX/Linuxとスクリプト言語	Linux基礎	初級	UNIXの基礎の理解	門田 幸二 (東京大学)	<a href="#">講義資料一覧 (PDF:32KB)</a>	
	13:15-14:45			中級			<a href="#">統合TV</a>	
	15:00-16:30							
	16:45-18:15							
7月24日 (金)	10:30-12:00		スクリプト言語	中級	シェルスクリプト	服部 恵美 (アメリエフ)	<a href="#">講義資料 (PDF:1.8MB)</a>	

# NGSハンズオン講習会

①事前予習事項のLinux部分は、日本乳酸菌学会誌に連載中のNGS解析記事をベースとしており、ウェブページ「**(Rで)塩基配列解析**」から全情報を取得可能

## 平成27年度NGSハンズオン講習会

平成27年度は、平成26年度の実績を踏まえ、講義内容の改善等を行い、ハンズオンに特化した、より効果的なNGS講習会を開催しました。

### H27年度概要

### H27年度講義日程・参考資料

H26年度講習会の情報については[こちら](#)をご覧ください。

### H27年度実施報告書・講義資料・動画等

● [講習会実施報告書 \(PDF: 2.17MB\)](#) および [受講者アンケート集計結果 \(データ集\) \(PDF: 662KB\)](#)

● [講義資料・動画](#) \*講義資料一覧のファイル名をクリックすると資料ファイル (PDF等) がダウンロードできます。

実施日	実施時間	大項目	項目	レベル	習得技術	担当講師(敬称略)	講義資料・動画(統合TV)
7月22日 (水)	10:30-12:00	PC環境の構築	Bio-Linux8とRのインストール状況確認		・Linux導入 ・R導入 ・NGS解析に必要な環境構築技術	門田 幸二 (東京大学)	<a href="#">事前予習資料一覧 (PDF:52KB)</a>
	13:15-14:45						
	15:00-16:30						<a href="#">講義資料一覧 (PDF:108KB)</a>
	16:45-18:15						
7月23日 (木)	10:30-12:00	UNIX/Linuxとスクリプト言語	Linux基礎	初級	UNIXの基礎の理解	門田 幸二 (東京大学)	<a href="#">講義資料一覧 (PDF:32KB)</a>
	13:15-14:45			中級			
	15:00-16:30						<a href="#">統合TV</a>
	16:45-18:15						
7月24日 (金)	10:30-12:00		スクリプト言語	中級	シェルスクリプト	服部 恵美 (アメリエフ)	<a href="#">講義資料 (PDF:1.8MB)</a>
	13:15-14:45						



# Contents

## ■ イン트로ダクション

- 概要、背景(NGS用カリキュラム、講習会)、Linuxスキル習得の意義
- ウェブ情報(日本乳酸菌学会誌のNGS連載やNGS講習会資料)

## ■ 実習環境に慣れる

- 仮想環境での作業に慣れる
- GUIとCUI(マウス操作かコマンド入力操作か)
- ターミナルでの作業
- 共有フォルダの概念を理解

## ■ 練習

- 作業ディレクトリの変更、練習用NGSデータファイルのダウンロード
- ファイルの確認、de novoゲノムアセンブリ
- BLAST検索

## ■ 課題

- グループごとに異なる課題ファイルを入力として、「ダウンロード、de novoアセンブリ、BLAST検索」を実行し、得られた結果をレポートにまとめて発表せよ
- グループ1はkadai1.fasta、グループ2はkadai2.fasta、etc.

# (Rで)塩基配列解析

## (Rで)塩基配列解析

(last modified 2016/08/12, since 2011)

このウェブページは [インストール](#) についての推奨手順 ([Windows2015.04.04版](#)と[Macintosh2015.04.03版](#)) に従ってフリーソフトRと必要なパッケージをインストール済みであるという前提で記述しています。初心者の方は [基本的な利用法](#) ([Windows2015.04.03版](#)と [Macintosh2015.04.03版](#)) で自習してください。本ウェブページを体系的にまとめた [書籍](#) もあります。(2015/04/03)

### What's new?

- [NGSハンズオン講習会2016](#)の第3部(8/4)の [講義資料PDF](#) を修正するところのミス、スライド140 (Bridgerに必要boostパッケージ実行もうまくいく)あたりです。(2016/08/12) **NEW**
- [NGSハンズオン講習会2016](#)の第2部(7/25-28)の講義資料修正
- [NGSハンズオン講習会2016](#)の第3部(8/01-04)の講義資料を一
- Twitterハッシュタグ [AJACS](#) はこちら。(2016/07/27) **NEW**
- [NGSハンズオン講習会2016](#)の第2部(7/25-28)の [コピペ用コマンド](#) データを生成する可能性があります。(2016/07/21) **NEW**

- 書籍 | [トランスクリプトーム解析 | 4.3.2 データの正規化\(応用編\)](#) (last modified 2014/04/27)
- 書籍 | [トランスクリプトーム解析 | 4.3.3 群間比較](#) (last modified 2014/04/28)
- 書籍 | [トランスクリプトーム解析 | 4.3.4 別の実験デザイン\(3群間\)](#) (last modified 2014/04/28)
- 書籍 | [日本乳酸菌学会誌 | 11月号](#) (last modified 2016/07/01)
- 書籍 | [日本乳酸菌学会誌 | 第1回イントロダクション](#) (last modified 2015/09/11)
- 書籍 | [日本乳酸菌学会誌 | 第2回GUI環境からコマンドライン環境へ](#) (last modified 2015/11/26)
- 書籍 | [日本乳酸菌学会誌 | 第3回Linux環境構築からNGSデータ取得まで](#) (last modified 2015/12/07)
- 書籍 | [日本乳酸菌学会誌 | 第4回クオリティコントロールとプログラムのインストール](#) (last modified 2016/01/07)
- 書籍 | [日本乳酸菌学会誌 | 第5回アセンブル、マッピング、そしてQC](#) (last modified 2016/07/05)
- 書籍 | [日本乳酸菌学会誌 | 第6回ゲノムアセンブリ](#) (last modified 2016/06/17)
- 書籍 | [日本乳酸菌学会誌 | 第7回ロングリードアセンブリ](#) (last modified 2016/07/01)
- イントロ | 一般 | [ランダムに行を抽出](#) (last modified 2014/07/17)
- イントロ | 一般 | [任意の文字列を行の最初に挿入](#) (last modified 2014/07/17)
- 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 2016/04/20)
- イントロ | 一般 | [ランダムな塩基配列を生成](#) (last modified 2014/06/16)

①②③で示す各回の原稿PDF (JSLAB\*\_kadota.pdf) は、デスクトップ上にあるhogeフォルダ内にあります

# 乳酸菌NGS連載

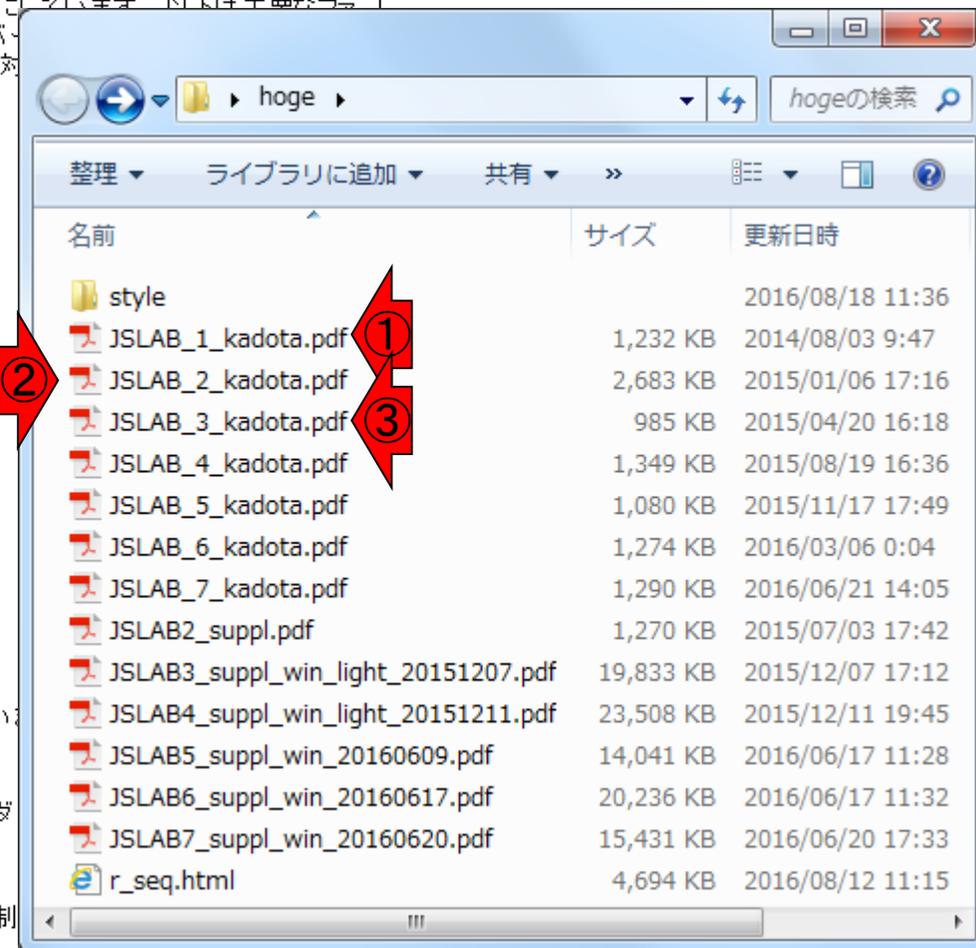
書籍 | 日本乳酸菌学会誌 | について **NEW**

(このウェブページの取扱い上、書籍としていますが学会誌です) [日本乳酸菌学会誌](#)の連載原稿を書いています。NGSデータ解析初心者用に、各種情報収集先、Linux環境構築、Linuxコマンドなど、講習会などに出なくても十分な学習効果が得られるような情報提供を目指して執筆しています。情報もできるだけWindows用とMacintosh用の両方を作成しています。原稿PDF、ウェブ資料を含めフリーでダウンロード可能です。本文中で触れたウェブサイトのリンク先などの情報も辿れるようになっています。以下は主要なファイルのみリストアップしています。ダウンロードしたPDFファイルのトップページ右上にある日付のバナーは、元のウェブページのキャッシュに残っているのが表示されてしまう現象に遭遇してしまっています。対応は「キャッシュを削除」です。

- [第1回イントロダウンロード](#) (2014年07月):
  - [原稿PDF](#) ①
- [第2回GUI環境からコマンドライン環境へ](#) (2014年11月):
  - [原稿PDF](#) ②
  - [ウェブ資料PDF](#) (2015.07.03版; 約2MB)
  - 1. VirtualBox, および2. Extension Packのインストール手順:
    - [Windows用](#) (2015.11.18版; 約3MB)
    - [Macintosh用](#) (2015.11.18版; 約8MB)
  - 3. 仮想マシンの作成、および4. Bio-Linux 8のisoファイルからのインストール手順:
    - [Windows用](#) (2015.11.19版; 約6MB)
    - [Macintosh用](#) (2015.11.19版; 約5MB)
  - Bio-Linux 8のovaファイルからのインストール手順(2015/11/25追加):
    - [Windows用](#) (2015.11.24版; 約2MB)
    - [Macintosh用](#) (2015.11.25版; 約4MB)
- [第3回Linux環境構築からNGSデータ取得まで](#) (2015年03月):

ユーザからの要望を踏まえ、ウェブ資料を更新しました。念のためオリジナル版も残してまいります(2015.12.07追加)。

  - [原稿PDF](#) ③
  - [ウェブ資料PDF](#) (オリジナル版; 原稿PDFと同じく、HDD150GB、約1.3億の全リードのデータ) (2015.07.01版; 約21MB; 非推奨)
    - [Windows用](#) (2015.07.01版; 約21MB; 非推奨)
    - [Macintosh用](#) (2015.04.27版; 約23MB)
  - [ウェブ資料PDF](#) (軽量版; 原稿PDFと異なり、HDD100GB、最初の150万リードのみに制限) (2015.12.07版; 約20MB; 推奨)
    - [Windows用](#) (2015.12.07版; 約20MB; 推奨)
    - [Macintosh用](#) (未着手)

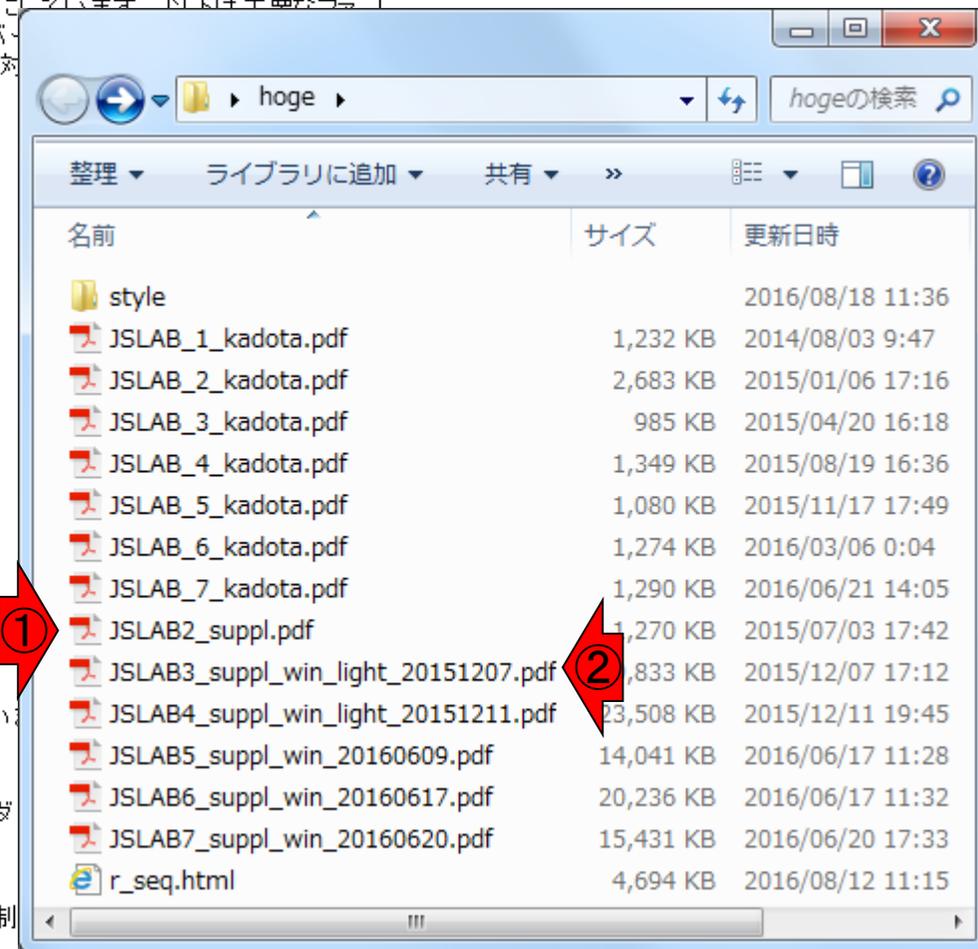


# 乳酸菌NGS連載

書籍 | 日本乳酸菌学会誌 | について **NEW**

(このウェブページの取扱い上、書籍としていますが学会誌です) [日本乳酸菌学会誌](#)の連載原稿を書いています。NGSデータ解析初心者用に、各種情報収集先、Linux環境構築、Linuxコマンドなど、講習会などに出なくても十分な学習効果が得られるような情報提供を目指して執筆しています。情報もできるだけWindows用とMacintosh用の両方を作成しています。原稿PDF、ウェブ資料を含めフリーでダウンロード可能です。本文中で触れたウェブサイトのリンク先などの情報も辿れるようにしています。以下は主要なファイルのみリストアップしています。ダウンロードしたPDFファイルのトップページ右上にある日付のバリエーションがあるウェブページのキャッシュに残っているのが表示されてしまう現象に遭遇してしまっています。対応は「キャッシュを削除」です。

- [第1回イントロダクション](#)(2014年07月):
  - [原稿PDF](#)
- [第2回GUI環境からコマンドライン環境へ](#)(2014年11月):
  - [原稿PDF](#)
  - [ウェブ資料PDF](#)(2015.07.03版; 約2MB)
    - ① 1. VirtualBox、および2. Extension Packのインストール手順:
      - [Windows用](#)(2015.11.18版; 約3MB)
      - [Macintosh用](#)(2015.11.18版; 約8MB)
    - 3. 仮想マシンの作成、および4. Bio-Linux 8のisoファイルからのインストール手順:
      - [Windows用](#)(2015.11.19版; 約6MB)
      - [Macintosh用](#)(2015.11.19版; 約5MB)
    - Bio-Linux 8のovaファイルからのインストール手順(2015/11/25追加):
      - [Windows用](#)(2015.11.24版; 約2MB)
      - [Macintosh用](#)(2015.11.25版; 約4MB)
- [第3回Linux環境構築からNGSデータ取得まで](#)(2015年03月):
 ユーザからの要望を踏まえ、ウェブ資料を更新しました。念のためオリジナル版も残してまいります。ご了承ください(2015.12.07追加)。
  - [原稿PDF](#)
  - [ウェブ資料PDF\(オリジナル版\)](#); 原稿PDFと同じく、HDD150GB、約1.3億の全リードのデータ
    - [Windows用](#)(2015.07.01版; 約21MB; 非推奨)
    - [Macintosh用](#)(2015.04.27版; 約23MB)
  - [ウェブ資料PDF\(軽量版\)](#); 原稿PDFと異なり、HDD100GB、最初の150万リードのみに制限
    - [Windows用](#)(2015.12.07版; 約20MB; 推奨) ②
    - [Macintosh用](#)(2015.12.07版; 約20MB; 推奨)



# 乳酸菌NGS連載

書籍 | 日本乳酸菌学会誌 | について **NEW**

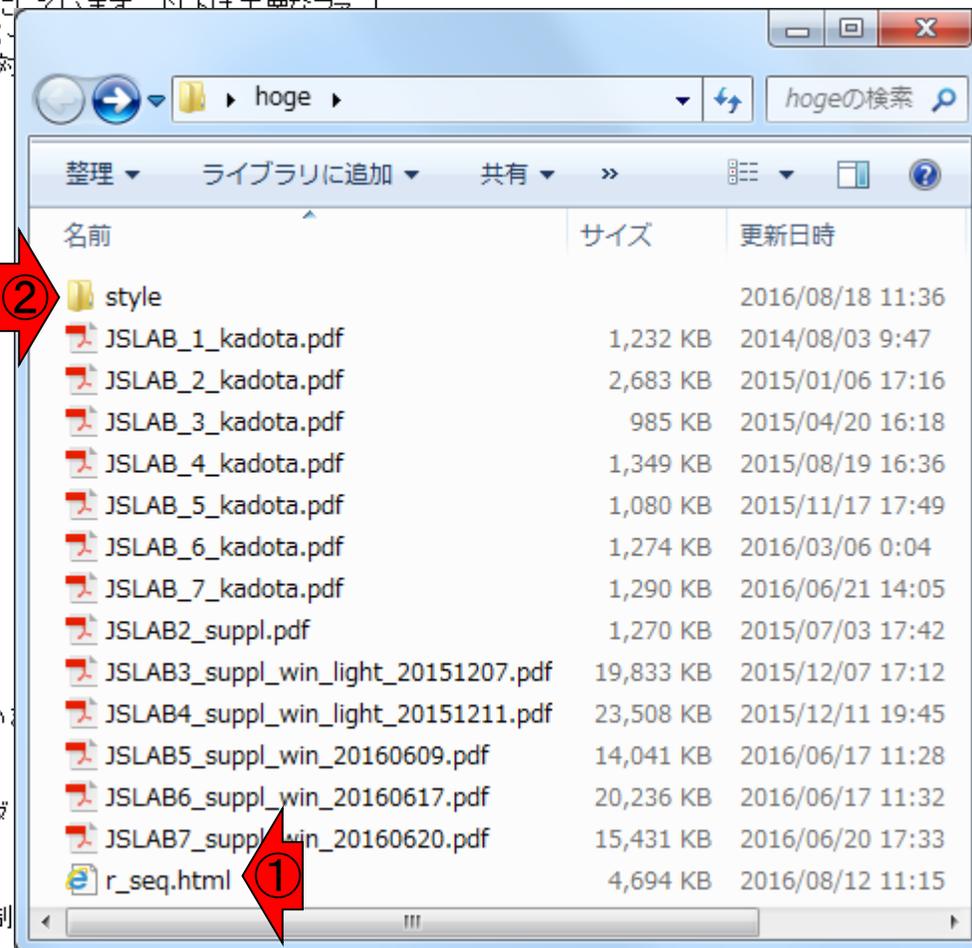
①は「(Rで)塩基配列解析」のソースファイル。ネットワーク不調時にダブルクリックで開くことで、ローカル環境でウェブページを開くことができます。②はウェブページの各種設定情報を含むフォルダです

(このウェブページの取扱い上、書籍としていますが学会誌です) [日本乳酸菌学会誌](#)の連載原稿を書いています。NGSデータ解析初心者用に、各種情報収集先、Linux環境構築、Linuxコマンドなど、講習会などに出なくても十分な学習効果が得られるような情報提供を目指して執筆しています。情報もできるだけWindows用とMacintosh用の両方を作成しています。原稿PDF、ウェブ資料を含めフリーでダウンロード可能です。本文中で触れたウェブサイトのリンク先などの情報も辿れるようにしています。以下は主要なファイルのみリストアップしています。ダウンロードしたPDFファイルのトップページ右上にある日付のバナーは、ブラウザのキャッシュに残っているのが表示されてしまう現象に遭遇してしまっています。対応として「キャッシュを削除」です。

- [第1回イントロダクション](#)(2014年07月):
  - [原稿PDF](#)
- [第2回GUI環境からコマンドライン環境へ](#)(2014年11月):
  - [原稿PDF](#)
  - [ウェブ資料PDF](#)(2015.07.03版; 約2MB)
  - 1. VirtualBox、および2. Extension Packのインストール手順:
    - [Windows用](#)(2015.11.18版; 約3MB)
    - [Macintosh用](#)(2015.11.18版; 約8MB)
  - 3. 仮想マシンの作成、および4. Bio-Linux 8のisoファイルからのインストール手順:
    - [Windows用](#)(2015.11.19版; 約6MB)
    - [Macintosh用](#)(2015.11.19版; 約5MB)
  - Bio-Linux 8のovaファイルからのインストール手順(2015/11/25追加):
    - [Windows用](#)(2015.11.24版; 約2MB)
    - [Macintosh用](#)(2015.11.25版; 約4MB)
- [第3回Linux環境構築からNGSデータ取得まで](#)(2015年03月):

ユーザからの要望を踏まえ、ウェブ資料を更新しました。念のためオリジナル版も残してまいります(2015.12.07追加)。

  - [原稿PDF](#)
  - ウェブ資料PDF(オリジナル版; 原稿PDFと同じく、HDD150GB、約1.3億の全リードのデータ)
    - [Windows用](#)(2015.07.01版; 約21MB; 非推奨)
    - [Macintosh用](#)(2015.04.27版; 約23MB)
  - ウェブ資料PDF(軽量版; 原稿PDFと異なり、HDD100GB、最初の150万リードのみに制限)
    - [Windows用](#)(2015.12.07版; 約20MB; 推奨)
    - [Macintosh用](#)(未着手)



# 連載第1回原稿

①乳酸菌NGS連載第1回原稿に相当する、②をダブルクリックで開いてみましょう

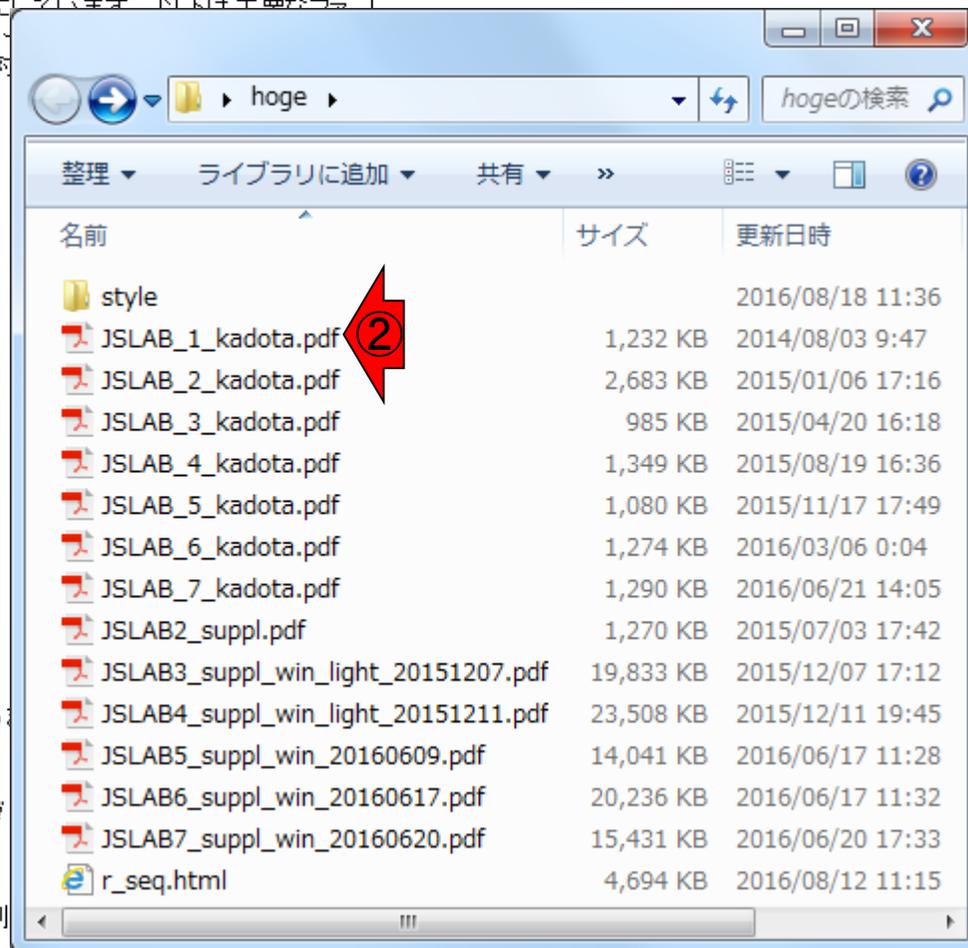
書籍 | 日本乳酸菌学会誌 | について **NEW**

(このウェブページの取扱い上、書籍としていますが学会誌です) [日本乳酸菌学会誌](#)の連載原稿を書いています。NGSデータ解析初心者用に、各種情報収集先、Linux環境構築、Linuxコマンドなど、講習会などに出なくても十分な学習効果が得られるような情報提供を目指して執筆しています。情報もできるだけWindows用とMacintosh用の両方を作成しています。原稿PDF、ウェブ資料を含めフリーでダウンロード可能です。本文中で触れたウェブサイトのリンク先などの情報も辿れるようにしています。以下は主要なファイルのみリストアップしています。ダウンロードしたPDFファイルのトップページ右上にある日付のバグ(ダウンロードしたPDFファイルのトップページ右上にある日付のバグ)のあるウェブページのキャッシュに残っているのが表示されてしまう現象に遭遇してしまっています。対応は「キャッシュを削除」です。

- [第1回イントロダウンロード](#)(2014年07月):
  - [原稿PDF](#) ①
- [第2回GUI環境からコマンドライン環境へ](#)(2014年11月):
  - [原稿PDF](#)
  - [ウェブ資料PDF](#)(2015.07.03版; 約2MB)
  - 1. VirtualBox、および2. Extension Packのインストール手順:
    - [Windows用](#)(2015.11.18版; 約3MB)
    - [Macintosh用](#)(2015.11.18版; 約8MB)
  - 3. 仮想マシンの作成、および4. Bio-Linux 8のisoファイルからのインストール手順:
    - [Windows用](#)(2015.11.19版; 約6MB)
    - [Macintosh用](#)(2015.11.19版; 約5MB)
  - Bio-Linux 8のovaファイルからのインストール手順(2015/11/25追加):
    - [Windows用](#)(2015.11.24版; 約2MB)
    - [Macintosh用](#)(2015.11.25版; 約4MB)
- [第3回Linux環境構築からNGSデータ取得まで](#)(2015年03月):

ユーザからの要望を踏まえ、ウェブ資料を更新しました。念のためオリジナル版も残してはいた  
用ください(2015.12.07追加)。

  - [原稿PDF](#)
  - [ウェブ資料PDF\(オリジナル版\)](#); 原稿PDFと同じく、HDD150GB、約1.3億の全リードのデータ  
    - [Windows用](#)(2015.07.01版; 約21MB; 非推奨)
    - [Macintosh用](#)(2015.04.27版; 約23MB)
  - [ウェブ資料PDF\(軽量版\)](#); 原稿PDFと異なり、HDD100GB、最初の150万リードのみに制限  
    - [Windows用](#)(2015.12.07版; 約20MB; 推奨)
    - [Macintosh用](#)(未着手)



# 連載第1回原稿

書籍 | 日本乳酸菌学会誌 | について

(このウェブページの取扱い上、書籍としていますが学  
初心者用に、各種情報収集先、Linux環境構築、Linux  
提供を目指して執筆しています。情報もできるだけWi  
めフリーでダウンロード可能です。本文中で触れたウェ  
イルのみリストアップしています。ダウンロードしたPD  
いるウェブページのキャッシュに残っているのが表示さ  
キャッシュを削除)です。

## ② 第1回イントロダクション(2014年07月):

・ [原稿PDF](#)

## ・ 第2回GUI環境からコマンドライン環境へ(2014年)

・ [原稿PDF](#)

・ [ウェブ資料PDF](#)(2015.07.03版; 約2MB)

・ 1. VirtualBox, および2. Extension Packの

▪ [Windows用](#)(2015.11.18版; 約3MB)

▪ [Macintosh用](#)(2015.11.18版; 約8MB)

・ 3. 仮想マシンの作成, および4. Bio-Linux

▪ [Windows用](#)(2015.11.19版; 約6MB)

▪ [Macintosh用](#)(2015.11.19版; 約5MB)

・ Bio-Linux 8のovaファイルからのインスト

▪ [Windows用](#)(2015.11.24版; 約2MB)

▪ [Macintosh用](#)(2015.11.25版; 約4MB)

## ・ 第3回Linux環境構築からNGSデータ取得まで(ユ ーザからの要望を踏まえ、ウェブ資料を更新 用ください(2015.12.07追加)。

・ [原稿PDF](#)

・ [ウェブ資料PDF\(オリジナル版; 原稿PDF\)](#)

▪ [Windows用](#)(2015.07.01版; 約21MB)

▪ [Macintosh用](#)(2015.04.27版; 約23MB)

・ [ウェブ資料PDF\(軽量版; 原稿PDFと異なる\)](#)

▪ [Windows用](#)(2015.12.07版; 約20MB)

▪ [Macintosh用\(未着手\)](#)

NGS解析周辺の講義内容を中心にま  
ンを行う。ウェブサイト (Rで) 塩基配列解析 (URL: [http://www.bi.a.u-tokyo.ac.jp/~kadota/1\\_seq.html](http://www.bi.a.u-tokyo.ac.jp/~kadota/1_seq.html)) 中に本連載で述べるリンク先を掲載してあるので効率的に活用してほしい。

Key words : NGS, RNA-seq, Bioinformatics, R, Linux

## 情報収集先 (講義、講習会、ウェブサイトなど)

筆者らの所属機関であるアグリバイオインフォマティクス教育研究プログラム (以下、アグリバイオ) では、数年前より NGS 関連ハンズオン講義 (ノート PC を用いた実習を含む講義のこと) の割合を徐々に高めている。大学院講義ではあるものの、東京大学以外の学生、一般企業の社会人、ポスドクも受講可能である。アグリバイオの主な目標は、高度なバイオインフォマティクス養成ではなく、手元にあるデータを自在に解析する技術を身につけたい実験系研究者の養成である。例年 20% 程度の受講生が東京大学以外であり、受講費用もかからないため、本誌読者向けの講義プログラムといえる。

\*To whom correspondence should be addressed.

Phone : +81-3-5841-2395

Fax : +81-3-5841-1136

E-mail : [kadota@bi.a.u-tokyo.ac.jp](mailto:kadota@bi.a.u-tokyo.ac.jp)

こんな感じのものが見えるはずですが。例えば原稿中の①「統合TV」のサイトはググってもよいが、各回のサイトからも辿れるようにしているのので、②第1回のサイトをクリック

NGS データ解析手法に関わるバイオインフォマティクスの多くは、日本バイオインフォマティクス学会 (JSBi) か NGS 現場の会に所属している。これらの年会や研究会への参加を通じた情報収集も有意義であろう。例えば、2014 年度の JSBi 年会は 10 月に仙台で開催されるが、いくつか講習会も企画されている。HPCI 人材養成プログラムも比較的規模の大きな活動を実施している。お台場にある産業技術総合研究所・CBRC を拠点として、NGS に特化した内容ではないものの、セミナー、ワークショップ、チュートリアル、e-learning などが積極的に実施されている。

e-learning 系の代表格としては統合 TV<sup>1)</sup> が挙げられる。文字通り様々なウェブツールやデータベースなどの使い方を紹介する番組である。“NGS” というキーワード検索でリストアップされるものは全 776 番組中 9 番組と意外に少ない (2014 年 5 月 4 日調べ) ものの、“次世代シーケンサ”で検索すると 20 番組程度がヒットする。また、バイオインフォマティクスの多くが利用している Linux (り

# 第1回のサイト

## 書籍 | 日本乳酸菌学会誌 | 第1

日本乳酸菌学会誌の第1回分です。

- [原稿PDF](#)

### 要約:

- [\(Rで\)塩基配列解析](#)はここ

### 情報収集先(講義、講習会、ウェブサイトなど):

- [アグリバイオインフォマティクス教育研究](#)
- [日本バイオインフォマティクス学会\(JSBi\)](#)
- [NGS現場の会](#)
- [産総研CBRC](#)
- [HPCI人材養成プログラム](#)
  - [e-learning](#)
  - [tutorial\(講習会\)](#)
  - [セミナー](#)
  - [ワークショップ](#)

② [統合TV: Kawano et al., Brief Bioinform.](#)

- [WindowsでUNIX! 1. Cygwin イン](#)
- [WindowsでUNIX! 2. ファイル操作](#)
- [WindowsでUNIX! 3. ファイル操作](#)
- [WindowsでUNIX! 4. ファイル操作](#)
- [Perlの使い方インストール編\(Wir](#)
- [Perlの使い方事例編\(前編\)](#)
- [Perlの使い方事例編\(後編\)](#)

NGS 解析周辺の講義内容を中心に基礎から応用まで幅広く述べる。第1回は、全体のイントロダクションを行う。ウェブサイト (Rで) 塩基配列解析 (URL: [http://www.iu.a.u-tokyo.ac.jp/~kadota/r\\_seq.html](http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html)) 中に本連載で述べるリンク先を掲載してあるので効率的に活用してほしい。

Key words : NGS, RNA-seq, Bioinformatics, R, Linux

### 情報収集先 (講義、講習会、ウェブサイトなど)

筆者らの所属機関であるアグリバイオインフォマティクス教育研究プログラム (以下、アグリバイオ) では、数年前より NGS 関連ハンズオン講義 (ノート PC を用いた実習を含む講義のこと) の割合を徐々に高めている。大学院講義ではあるものの、東京大学以外の学生、一般企業の社会人、ポスドクも受講可能である。アグリバイオの主な目標は、高度なバイオインフォマティクス養成ではなく、手元にあるデータを自在に解析する技術を身につけたい実験系研究者の養成である。例年 20% 程度の受講生が東京大学以外であり、受講費用もかからないため、本誌読者向けの講義プログラムといえる。

\*To whom correspondence should be addressed.

Phone : +81-3-5841-2395  
Fax : +81-3-5841-1136  
E-mail : [kadota@bi.a.u-tokyo.ac.jp](mailto:kadota@bi.a.u-tokyo.ac.jp)

NGS データ解析手法に関わるバイオインフォマティクスの多くは、日本バイオインフォマティクス学会 (JSBi) か NGS 現場の会に所属している。これらの年会や研究会への参加を通じた情報収集も有意義であろう。例えば、2014 年度の JSBi 年会は 10 月に仙台で開催されるが、いくつか講習会も企画されている。HPCI 人材養成プログラムも比較的規模の大きな活動を実施している。お台場にある産業技術総合研究所・CBRC を拠点として、NGS に特化した内容ではないものの、セミナー、ワークショップ、チュートリアル、e-learning などが積極的に実施されている。

e-learning 系の代表格としては統合 TV<sup>1)</sup> が挙げられる。文字通り様々なウェブツールやデータベースなどの使い方を紹介する番組である。“NGS” というキーワード検索でリストアップされるものは全 776 番組中 9 番組と意外に少ない (2014 年 5 月 4 日調べ) もの、“次世代シーケンサ” で検索すると 20 番組程度がヒットする。また、バイオインフォマティクスの多くが利用している Linux (り

# 第1回のサイト

書籍 | 日本乳酸菌学会誌 | 第1回イントロダクション

日本乳酸菌学会誌の第1回分です。

• [原稿PDF](#)

要約:

• [\(Rで\)塩基配列解析](#)はここ

情報収集先(講義、講習会、ウェブサイトなど):

• [アグリバイオインフォマティクス教育研究プログラム](#)

• [日本バイオインフォマティクス学会\(JSBi\)](#)

• [NGS現場の会](#)

• [産総研CBRC](#)

• [HPCI人材養成プログラム](#)

◦ [e-learning](#)

◦ [tutorial\(講習会\)](#)

◦ [セミナー](#)

◦ [ワークショップ](#)

• [統合TV: Kawano et al., Brief Bioinform., 2012](#)

◦ [WindowsでUNIX! 1. Cygwin インストール編](#)

◦ [WindowsでUNIX! 2. ファイル操作編](#)

◦ [WindowsでUNIX! 3. ファイル操作応用編](#)

◦ [WindowsでUNIX! 4. ファイル操作発展編](#)

◦ [Perlの使い方 インストール編\(Windows\)](#)

◦ [Perlの使い方 事例編\(前編\)](#)

◦ [Perlの使い方 事例編\(後編\)](#)

各回のウェブサイトを用意することで、統合TVの①原著論文へのリンクや、②統合TVで提供している具体的な番組名やそのリンク先を示すことができます。また、③ページ下部に移動して眺めると、提供している情報量も膨大であることがわかります

③

①

②

# 第1回のサイト

## 書籍 | 日本乳酸菌学会誌 | 第1回イントロダクション

日本乳酸菌学会誌の第1回分です。

- [原稿PDF](#)

要約:

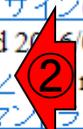
- [\(Rで\)塩基配列解析](#)はここ

情報収集先(講義、講習会、ウェブサイトなど):

- [アグリバイオインフォマティクス教育研究プログラム](#)
- [日本バイオインフォマティクス学会\(JSBi\)](#)
- [NGS現場の会](#)
- [産総研CBRC](#)
- [HPCI人材養成プログラム](#)
  - [e-learning](#)
  - [tutorial\(講習会\)](#)
  - [セミナー](#)
  - [ワークショップ](#)
- [統合TV: Kawano et al., Brief Bioinform., 2012](#)
  - [WindowsでUNIX! 1. Cygwin インストール編](#)
  - [WindowsでUNIX! 2. ファイル操作編](#)
  - [WindowsでUNIX! 3. ファイル操作応用編](#)
  - [WindowsでUNIX! 4. ファイル操作発展編](#)
  - [Perlの使い方 インストール編\(Windows\)](#)
  - [Perlの使い方 事例編\(前編\)](#)
  - [Perlの使い方 事例編\(後編\)](#)



- 書籍 | トランスクリプトーム解析 | [4.3.2 データの正規化\(応用編\)](#) (last modified 2014/04/27)
- 書籍 | トランスクリプトーム解析 | [4.3.3 2群間比較](#) (last modified 2014/04/28)
- 書籍 | トランスクリプトーム解析 | [4.3.4 他の実験デザイン\(3群間\)](#) (last modified 2014/04/28)
- 書籍 | [日本乳酸菌学会誌](#) | [第1回イントロダクション](#) (last modified 2015/07/01)
- 書籍 | [日本乳酸菌学会誌](#) | [第2回GUI環境からコマンドライン環境へ](#) (last modified 2015/11/26)
- 書籍 | [日本乳酸菌学会誌](#) | [第3回Linux環境構築からNGSデータ取得まで](#) (last modified 2015/12/07)
- 書籍 | [日本乳酸菌学会誌](#) | [第4回クオリティコントロールとプログラムのインストール](#) (last modified 2016/07/01)
- 書籍 | [日本乳酸菌学会誌](#) | [第5回アセンブル、マッピング、そしてQC](#) (last modified 2016/07/05)
- 書籍 | [日本乳酸菌学会誌](#) | [第6回ゲノムアセンブリ](#) (last modified 2016/06/17)
- 書籍 | [日本乳酸菌学会誌](#) | [第7回ロングリードアセンブリ](#) (last modified 2016/07/01)
- インタロ | 一般 | [ランダムに行を抽出](#) (last modified 2014/07/17)
- インタロ | 一般 | [任意の文字列を行の最初に挿入](#) (last modified 2014/07/17)
- インタロ | 一般 | [任意のキーワードを含む行を抽出\(基礎\)](#) (last modified 2016/04/20)
- インタロ | 一般 | [ランダムな塩基配列を生成](#) (last modified 2014/06/16)



# Tips

ウェブブラウザのサイズを変更したりすると、自分がどこにいるのかよくわからなくなります。その場合は、常に右下部分に見えている①「トップページへ」をクリックして…

The screenshot shows a web browser window with the address bar containing the URL `http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#`. The browser title is "(Rで)塩基配列解析". The main content area displays R code for plotting a grid and legend. Below the code, there is a section header "書籍 | トランスクリプトーム解析 | 4.3.1 シミュレーションデータ(負の二項分布)". The text below the header discusses the source of the R code and provides instructions for copying it. A code block contains `set.seed(100)`. Another section, "p189-190の網掛け部分:", shows a code block with `out_f <- "hypoData1.txt"` and a comment "#出力ファイル名を指定してout\_fに格納". A red arrow with the number 1 points to the text "トップページへ" in the comment.

```
xlim=c(1e-01, 1e+07), ylim=c(1e-01, 1e+07), col="gray")
### Legendなど ###
grid(col="gray", lty="dotted")
abline(a=0, b=1, col="gray")
legend("bottomright",c("male+female", "female only"), col=c("black", "gray"), pch=20
```

## 書籍 | トランスクリプトーム解析 | 4.3.1 シミュレーションデータ(負の二項分布)

シリーズ [Useful R 第7巻トランスクリプトーム解析](#)のp188-196のRコードです。  
「ファイル」-「ディレクトリの変更」でデスクトップに移動し以下をコピー。

**p189:**

「`1setwd("C:/Users/kadota/Desktop/")`と`setwd`関数の前に1が入っていますが間違いです`m(_ )m`。原稿校正時に混入したものと思われます。

```
set.seed(100)
```

**p189-190の網掛け部分:**

```
out_f <- "hypoData1.txt"           #出力ファイル名を指定してout_fに格納
### シミュレーションデータの作成 ###
```

① [トップページへ](#)

# Tips

## (Rで)塩基配列解析

(last modified 2016/08/12, since 2011)



このウェブページは [インストール](#) についての推奨手順 ([Windows2015.04.04版](#)と[Macintosh2015.04.03版](#))に従ってフリーソフトRと必要なパッケージをインストール済みであるという前提で記述しています。初の方は [基本的な利用法](#) ([Windows2015.04.03版](#)と [Macintosh2015.04.03版](#))で自習してください。本ウェブページを体系的にまとめた [書籍](#) もあります。(2015/04/03)

ウェブブラウザのサイズを変更したりすると、自分がどこにいるのかよくわからなくなります。その場合は、常に右下部分に見えている①「トップページへ」をクリックして、②「(Rで)塩基配列解析」のタイトルが見える一番上まで移動したのち、例えば③NGS連載第2回のページをクリックするなどすればよい

### What's new?

- [NGSハンズオン講習会2016](#)の第3部(8/4)の [講義資料PDF](#) を修正するところのミス、スライド140 (Bridger)に必要なboostパッケージ実行もうまくいくあたりです。(2016/08/12) **NEW**
- [NGSハンズオン講習会2016](#)の第2部(7/25-28)の講義資料修正
- [NGSハンズオン講習会2016](#)の第3部(8/01-04)の講義資料を一覧
- Twitterハッシュタグ [AJACS](#) はこちら。(2016/07/27) **NEW**
- [NGSハンズオン講習会2016](#)の第2部(7/25-28)の [コピペ用コマンド](#) データを生成する可能性があります。(2016/07/21) **NEW**

- 書籍 | [トランスクリプトーム解析 | 4.3.2 データの正規化\(応用編\)](#) (last modified 2014/04/27)
- 書籍 | [トランスクリプトーム解析 | 4.3.3 群間比較](#) (last modified 2014/04/28)
- 書籍 | [トランスクリプトーム解析 | 4.3.4 他の実験デザイン\(3群間\)](#) (last modified 2014/04/28)
- 書籍 | [日本乳酸菌学会誌 | について](#) (last modified 2016/07/01)
- 書籍 | [日本乳酸菌学会誌 | 第1回イントロダクション](#) (last modified 2015/09/11)
- 書籍 | [日本乳酸菌学会誌 | 第2回GUI環境からコマンドライン環境へ](#) (last modified 2015/11/26)
- 書籍 | [日本乳酸菌学会誌 | 第3回Linux環境構築からNGSデータ取得まで](#) (last modified 2015/12/07)
- 書籍 | [日本乳酸菌学会誌 | 第4回クオリティコントロールとプログラムのインストール](#) (last modified 2016/06/17)
- 書籍 | [日本乳酸菌学会誌 | 第5回アセンブル、マッピング、そしてQC](#) (last modified 2016/07/05)
- 書籍 | [日本乳酸菌学会誌 | 第6回ゲノムアセンブリ](#) (last modified 2016/06/17)
- 書籍 | [日本乳酸菌学会誌 | 第7回ロングリードアセンブリ](#) (last modified 2016/07/01)
- イントロ | 一般 | [ランダムに行を抽出](#) (last modified 2014/07/17)
- イントロ | 一般 | [任意の文字列を行の最初に挿入](#) (last modified 2014/07/17)
- 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 2016/04/20)
- イントロ | 一般 | [ランダムな塩基配列を生成](#) (last modified 2014/06/16)



# Tips

あるいは、「①CTRL + ②F」キーを押して、③「コマンドライン」などの任意のキーワードを入力し、ページ内検索をしてもよい。1つの項目中で示されている情報量が膨大なため、実際問題としてこのサイト利用時にはキーワード検索もよく用いる

検索: コマンドライン ③

- 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) | [NGSハンズオン講習会2016](#) (last modified 2016/08/12) **NEW**
- 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) (last modified 2015/11/13)
- 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) (last modified 2015/02/11)
- [書籍 | トランスクリプトーム解析 | について](#) (last modified 2014/0/)
- [書籍 | トランスクリプトーム解析 | 2.3.1 RNA-seqデータ\(FASTQ\)](#)
- [書籍 | トランスクリプトーム解析 | 2.3.2 リファレンス配列](#) (last modified 2014/0/)
- [書籍 | トランスクリプトーム解析 | 2.3.3 アンテーション情報](#) (last modified 2014/0/)
- [書籍 | トランスクリプトーム解析 | 2.3.4 マッピング\(準備\)](#) (last modified 2014/0/)
- [書籍 | トランスクリプトーム解析 | 2.3.5 マッピング\(本番\)](#) (last modified 2014/0/)
- [書籍 | トランスクリプトーム解析 | 2.3.6 カウントデータ取得](#) (last modified 2014/0/)
- [書籍 | トランスクリプトーム解析 | 3.3.1 解析目的別留意点](#) (last modified 2014/0/)
- [書籍 | トランスクリプトーム解析 | 3.3.2 データの正規化\(基礎編\)](#)
- [書籍 | トランスクリプトーム解析 | 3.3.3 クラスターリング](#) (last modified 2014/0/)
- [書籍 | トランスクリプトーム解析 | 3.3.4 各種プロット](#) (last modified 2014/0/)
- [書籍 | トランスクリプトーム解析 | 4.3.1 シミュレーションデータ\(負\)](#)
- [書籍 | トランスクリプトーム解析 | 4.3.2 データの正規化\(応用編\)](#)
- [書籍 | トランスクリプトーム解析 | 4.3.3 2群間比較](#) (last modified 2014/0/)
- [書籍 | トランスクリプトーム解析 | 4.3.4 他の実験デザイン\(3群間\)](#)
- [書籍 | 日本乳酸菌学会誌 | について](#) (last modified 2016/07/01)
- [書籍 | 日本乳酸菌学会誌 | 第1回イントロダクション](#) (last modified 2016/08/18) **NEW**
- [書籍 | 日本乳酸菌学会誌 | 第2回GUI環境からコマンドライン環境へ](#) (last modified 2015/11/26)
- [書籍 | 日本乳酸菌学会誌 | 第3回Linux環境構築からNGSデータ取得まで](#) (last modified 2015/11/26) **トップページへ**
- [書籍 | 日本乳酸菌学会誌 | 第4回クオリティコントロールとプログラムのインストール](#) (last modified 2016/06/03)



# NGSハンズオン講習会

①H28年度の講習会(のLinux部分)は、②乳酸菌NGS連載第1-4回を予習として課した。予習事項は大まかに「仮想環境構築、Bio-Linux上での作業、共有フォルダやLinux系用語に慣れる、Linuxコマンドを一通り習得、...」

## バイオインフォマティクス人材育成のための講習会

### NGS解析

ライフサイエンス分野の研究現場においては、莫大・多様な研究データが産出され、取り扱うデータ量が飛躍的に増えています。一方で、それらのデータを整備・活用するための人材は不足している状況です。そのような研究現場の状況を踏まえ、NBDC運営委員会人材育成分科会において、研究データを整備・活用するバイオインフォマティクス人材を育成するためのカリキュラムに基づき実施した講習会です。本カリキュラムは、対応が急務であると思われる次世代シーケンサデータに焦点をあてた内容となっております。

#### ●人材育成分科会で策定したカリキュラム

[バイオインフォマティクス人材育成カリキュラム \(次世代シーケン\)](#)

[カリキュラムで習得できる技能](#)

[カリキュラム フロー図](#)

#### ●H28年度NGSハンズオン講習会 (2016年7月19日~8月4日)

※受講者アンケートにご協力をお願いします (予めメールでお知らせ)

[受講者アンケート回答はこちら](#)

回答期間: 7/19 (火) ~9/16 (金)

- 書籍 | トランスクリプトーム解析 | [4.3.2 データの正規化\(応用編\)](#) (last modified 2014/04/27)
- 書籍 | トランスクリプトーム解析 | [4.3.3 2群間比較](#) (last modified 2014/04/28)
- 書籍 | トランスクリプトーム解析 | [4.3.4 他の実験デザイン\(3群間\)](#) (last modified 2014/04/28)
- 書籍 | [日本乳酸菌学会誌](#) | [について](#) (last modified 2016/07/01)
- 書籍 | [日本乳酸菌学会誌](#) | [第1回イントロダクション](#) (last modified 2015/09/11)
- 書籍 | [日本乳酸菌学会誌](#) | [第2回GUI環境からコマンドライン環境へ](#) (last modified 2015/12/26)
- 書籍 | [日本乳酸菌学会誌](#) | [第3回Linux環境構築からNGSデータ取得まで](#) (last modified 2015/12/07)
- 書籍 | [日本乳酸菌学会誌](#) | [第4回クオリティコントロールとプログラムのインストール](#) (last modified 2015/12/07)
- 書籍 | [日本乳酸菌学会誌](#) | [第5回アセンブル、マッピング、そしてQC](#) (last modified 2016/07/05)
- 書籍 | [日本乳酸菌学会誌](#) | [第6回ゲノムアセンブリ](#) (last modified 2016/06/17)
- 書籍 | [日本乳酸菌学会誌](#) | [第7回ロングリードアセンブリ](#) (last modified 2016/07/01)
- インタロ | 一般 | [ランダムに行を抽出](#) (last modified 2014/07/17)
- インタロ | 一般 | [任意の文字列を行の最初に挿入](#) (last modified 2014/07/17)
- インタロ | 一般 | [任意のキーワードを含む行を抽出\(基礎\)](#) (last modified 2016/04/20)
- インタロ | 一般 | [ランダムな塩基配列を生成](#) (last modified 2014/06/16)

#### ●H27年度NGSハンズオン講習会 (2015年7月22日~8月6日)

#### ●H26年度NGS速習コース講習会 (2014年9月1日~12日)

# NGSハンズオン講習会

乳酸菌NGS連載第1-4回の予習事項をマスターしておけば、後は①で公開されている講義資料や動画(統合TVで今年度中に公開予定)で独習可能。**時代はe-learningでハンズオン**

## H28年度 NGSハンズオン講習会カリキュラム

### H28年度日程・講義資料・動画等

\* 講義資料・動画については、順次公開いたします(2016.8.18更新)。

カリキュラム (PDF: 2KB)

①

実施日	実施時間	大項目	タイトル	内容(予定)	担当講師 (敬称略)	講義資料・ 動画(統合TV)
7月19日 (火)	10:30- 18:15	はじめに(講習会参加者必読)  PC環境の構築	Bio-Linux8とRのインストール状況確認	<ul style="list-style-type: none"> <li>・ Bio-Linux8 (第2部および3部で利用するovaファイル)の導入確認</li> <li>・ 共有フォルダ設定完了確認</li> <li>・ 基本的なLinuxコマンドの習得状況確認</li> <li>・ R本体およびパッケージのインストール確認</li> <li>・ 講師指定の事前予習内容の再確認</li> <li>・ 講習会期間中に貸与されるノートPCを用いた各種動作確認</li> </ul>	主催・共催機関	講義資料 (PDF:4MB)
7月20日 (水)	10:30- 18:15	第1部 統計解析 (農学生命情報科学特論I)	ゲノム解析、塩基配列解析	<ul style="list-style-type: none"> <li>・ NGS解析手段、ウェブツール(DDBJ Pipeline)との連携</li> <li>・ k-mer解析(k個の連続塩基に基づく各種解析)の基礎と応用</li> <li>・ 塩基ごとの出現頻度解析(k=1)、2連続塩基の出現頻度解析(k=2)</li> <li>・ 塩基配列解析を行うための基本スキルの復習や作図</li> <li>・ de novoアセンブリ時のエラー補正やゲノムサイズ推定の基本的な考え方</li> </ul>	門田 幸二 (東京大学)	講義資料 (PDF:7.3MB)  解析データ (ZIP:2.2MB)
7月21日	10:30-		トランスクリプトーム	・ カウンティング取得以降の統計解析(RNA-seq)		講義資料

# ちなみに



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

Agricultural Bioinformatics Research Unit

+ サイトマップ + Eng



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



アグリバイオ単体で行う大学院講義では、Linux環境でのデータ解析系講義は行われません(と思っておけば間違いありません)。受講人数が多すぎる(①最大で130名)、受講生の意識レベルや習熟度の差が大きく、講義として成立させることが困難なためです

ようこそ!!  
アグリバイオインフォマティクス  
教育研究ユニットへ

バイオインフォマティクスの実践的基礎教育から、関連した農学生命科学の教育と研究指導、本分野の社会連携、国際拠点の形成を目指して



### お知らせ - 受講に関する更新情報

- ▶ アグリバイオの講義を受講するには「**2016年度受講証**」が必要となります。受講証に記載されている**バーコード**がパソコンの貸出と出席登録に必要となり講義の際には、受講証を必ず持参して下さい。受講証は受講登録手続後、農学部2号館地下1階14-2号室で発行されます。
- ▶ 平成28年度以前に受講された方は、**受講IDの修正処理**をしてください。誤りますと、過去に取得した単位が加算されません。修正処理は、受講登録期間中でのみ行うことができます。
- ▶ 平成28年度受講受講生募集要項は [こちら](#)(PDF)です。
- ▶ 受講に関しての質問はまず[こちら](#)をごらんください。  
[Q & A集\(本学の大学院生の方\)](#)  
[Q & A集\(本学の大学院生以外の方\)](#)
- ▶ 成績証明書の発行を希望される方は申込用紙「**Word形式、PDF形式**」に必要事項を記入し、事務局までご連絡ください。



# 本講義では...

門田担当分は、(主にNGS解析を意識した)  
**Linuxスキルの習得が主目的**。何かをやった  
という達成感も得られるように、実際のNGSデ  
ータの一部を用いてゲノムアセンブリまで行う

## ■ キーワード

- NGS, Linux, バイオインフォマティクス, 仮想環境, Bio-Linux, ゲノムアセンブリ

## ■ Linux

- WindowsやMacintoshと同じく、OSの一種
- バイオインフォマティクス分野でよく利用される
- 「Windowsのコマンドプロンプト」や「Macintoshのターミナル」と同じく、lsやcdなどのLinuxコマンドを知らなければ何もできないため、慣れるまでが大変
- 使いこなせれば、最先端の解析用プログラムを自在にインストール・利用可能となり、効率的かつ**通り一辺倒でない**データ解析も可能となる

## ■ 次世代シーケンサ(NGS)解析

- NGSとは、大量の塩基配列を出力する実験機器(Next-Generation Sequencer)またはその技術を指す。主にゲノム解析やトランスクリプトーム解析と呼ばれる分野で利用されている
- 塩基配列解析用プログラムは、UNIX(今のLinux)環境で動作するものが多かった歴史的背景などから、現在でもLinux上で動くプログラムがまず最初に開発される場合が多い

# 本講義では...

実際に行うのは①の一部。それでも実際に手を動かし門田提供教材のノリに慣れておくことで、②の自習にもつながる。②の講習会やアグリバイオ大学院講義は、日本最大の受講人数規模(東大生以外の学生、社会人、ポスドク、教員なども含む)。ここで紹介したやり方をベースにすれば、情報共有もやりやすいと思われます

## バイオインフォマティクス人材育成のための講習会

### NGS解析

ライフサイエンス分野の研究現場においては、莫大・多様な研究データが産出され、取り扱うデータ量が飛躍的に増えています。一方で、それらのデータを整備・活用するための人材は不足している状況です。そのような研究現場の状況を踏まえ、NBDC運営委員会人材育成分科会において、研究データを整備・活用するバイオインフォマティクス人材を育成するためのカリキュラムに基づき実施した講習会です。本カリキュラムは、対応が急務であると思われる次世代シーケンサデータに焦点をあてた内容となっております。

#### ●人材育成分科会で策定したカリキュラム

[バイオインフォマティクス人材育成カリキュラム \(次世代シーケン\)](#)

カリキュラムで習得できる技能

カリキュラム フロー図

#### ●H28年度NGSハンズオン講習会 (2016年7月19日~8月4日)

※受講者アンケートにご協力をお願いいたします (予めメールでお知らせ)

受講者アンケート回答はこちら

回答期間: 7/19 (火) ~9/16 (金)

#### ●H27年度NGSハンズオン講習会 (2015年7月22日~8月6日)

#### ●H26年度NGS速習コース講習会 (2014年9月1日~12日)

- 書籍 | トランスクリプトーム解析 | [4.3.2 データの正規化\(応用編\)](#) (last modified 2014/04/27)
- 書籍 | トランスクリプトーム解析 | [4.3.3 2群間比較](#) (last modified 2014/04/28)
- 書籍 | トランスクリプトーム解析 | [4.3.4 他の実験デザイン\(3群間\)](#) (last modified 2014/04/28)
- 書籍 | [日本乳酸菌学会誌](#) | [について](#) (last modified 2016/07/01)
- 書籍 | [日本乳酸菌学会誌](#) | [第1回イントロダクション](#) (last modified 2015/09/11)
- 書籍 | [日本乳酸菌学会誌](#) | [第2回GUI環境からコマンドライン環境へ](#) (last modified 2015/11/26)
- 書籍 | [日本乳酸菌学会誌](#) | [第3回Linux環境構築からNGSデータ取得まで](#) (last modified 2015/12/07)
- 書籍 | [日本乳酸菌学会誌](#) | [第4回クオリティコントロールとプログラムのインストール](#) (last modified 2016/07/05)
- 書籍 | [日本乳酸菌学会誌](#) | [第5回アセンブル、マッピング、そしてQC](#) (last modified 2016/07/05)
- 書籍 | [日本乳酸菌学会誌](#) | [第6回ゲノムアセンブリ](#) (last modified 2016/06/17)
- 書籍 | [日本乳酸菌学会誌](#) | [第7回ロングリードアセンブリ](#) (last modified 2016/07/01)
- イントロ | 一般 | [ランダムに行を抽出](#) (last modified 2014/07/17)
- イントロ | 一般 | [任意の文字列を行の最初に挿入](#) (last modified 2014/07/17)
- 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 2016/04/20)
- イントロ | 一般 | [ランダムな塩基配列を生成](#) (last modified 2014/06/16)

# Contents

## ■ イントロダクション

- 概要、背景(NGS用カリキュラム、講習会)、Linuxスキル習得の意義
- ウェブ情報(日本乳酸菌学会誌のNGS連載やNGS講習会資料)

## ■ 実習環境に慣れる

- 仮想環境での作業に慣れる
- GUIとCUI(マウス操作かコマンド入力操作か)
- ターミナルでの作業
- 共有フォルダの概念を理解

## ■ 練習

- 作業ディレクトリの変更、練習用NGSデータファイルのダウンロード
- ファイルの確認、de novoゲノムアセンブリ
- BLAST検索

## ■ 課題

- グループごとに異なる課題ファイルを入力として、「ダウンロード、de novoアセンブリ、BLAST検索」を実行し、得られた結果をレポートにまとめて発表せよ
- グループ1はkadai1.fasta、グループ2はkadai2.fasta、etc.

# VirtualBoxを起動

## データ解析環境 (Linux)

NGS データ解析を高速かつ効率的に実行するプログラムの多くは、Linux というデータ解析環境上でのみ動作する。それゆえ、バイオインフォマティクスを目指す場合には、しばしば Linux 環境構築が最初の課題として与えられる。最近では、普段利用している Windows または Macintosh マシン上で Linux 環境を同時に利用するための手軽な手段も提供されている。この種の勉強を本格的に始めようとする、慣れない用語に戸惑う読者は意外に多い。まず Linux とは何か？ Linux は、Windows や Mac OS などと同じオペレーティングシステム (以下、OS) の一種である。多少の誤解を恐れずに説明する。Windows や Macintosh が安定して使いやすい大衆車だとすると、Linux は乗りこなせると速いスポーツカーのようなものである。次に戸惑うのは Linux の種類 (ディストリビューション) に関する記述である。Windows に Windows 7 や Windows 8.1 があるように、Linux にも様々な種類がある。Linux (スポーツカー) に関する情報収集をしているつもりでも、いつのまにか説明内容が Ubuntu (うぶんつ、と読む) とか LinuxMint とか CentOS などに切り替わっていてだんだん混乱してくる場合が多いだろう。Ubuntu と

連載第1回原稿のp88。Windows PC上でLinuxを動かす際に、①仮想化ソフトを利用。貸与PCには②VirtualBoxという仮想化ソフトがインストールされています。デスクトップ上にある③のアイコンをダブルクリックで起動

一種のコントローラと解釈してもよいだろう。通常は1台の車しか運転できないが、仮想化ソフトというコントローラのおかげで2台の車を同時に操作することができる。しばしば動作が不安定になるのは2台の車を同時に操作するという複雑なことをしているためであると解釈すればよい。

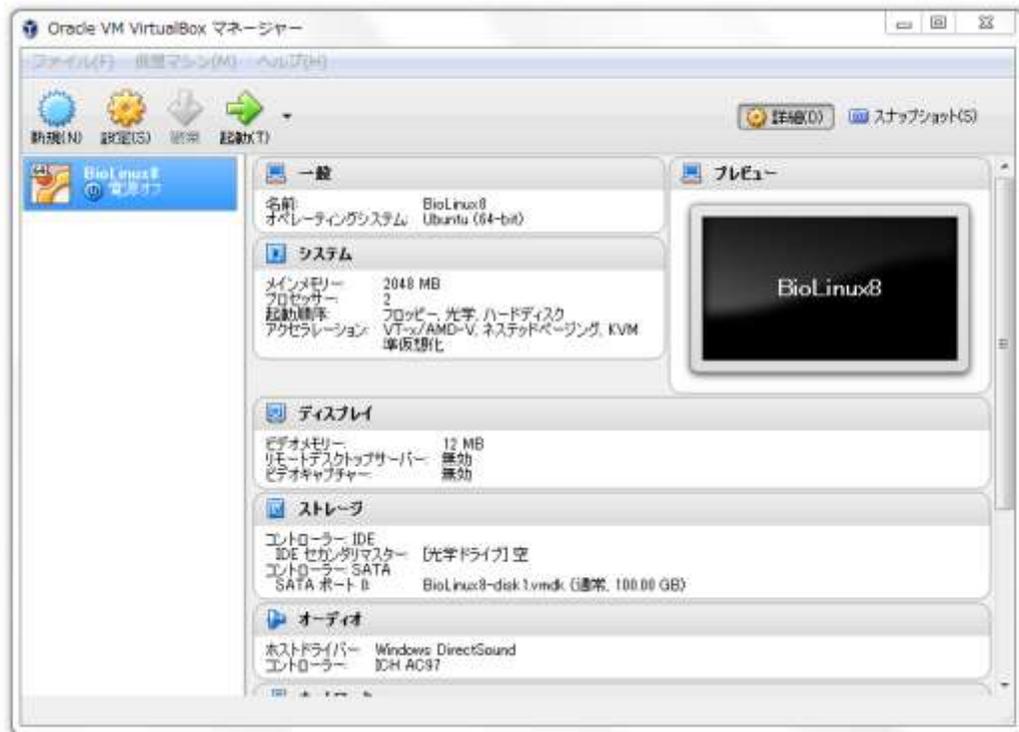
代表的な無料の仮想化ソフトとしては、VMware 社の VMware Player とオラクル社の VirtualBox の2つが挙げられる。VMware Player は非営利に限り<sup>①</sup>であり、歴史が古く安定している。VirtualBox は比較的最近開発されたソフトウェアで利便性が高い一方、やや動作が不安定だという声をきく。アグリバイオの講義で使用する PC には VMware Player をインストールしている。理由は、ウェブ上で収集可能な情報量が多くトラブル対応が比較的容易だったからである (図1)。

アSEMBL など多くの NGS 解析手法を自在に操りたいバイオインフォマティクス中級～上級を目指したい場合には、是非 Linux 環境構築に挑戦してほしい。最初は手持ちのノート PC で十分である。cd, ls, grep などの独特なコマンド操作、見慣れない GUI 画面、Windows 向けのソフトウェアなどとは異なりダブルクリックでプログラムのイ

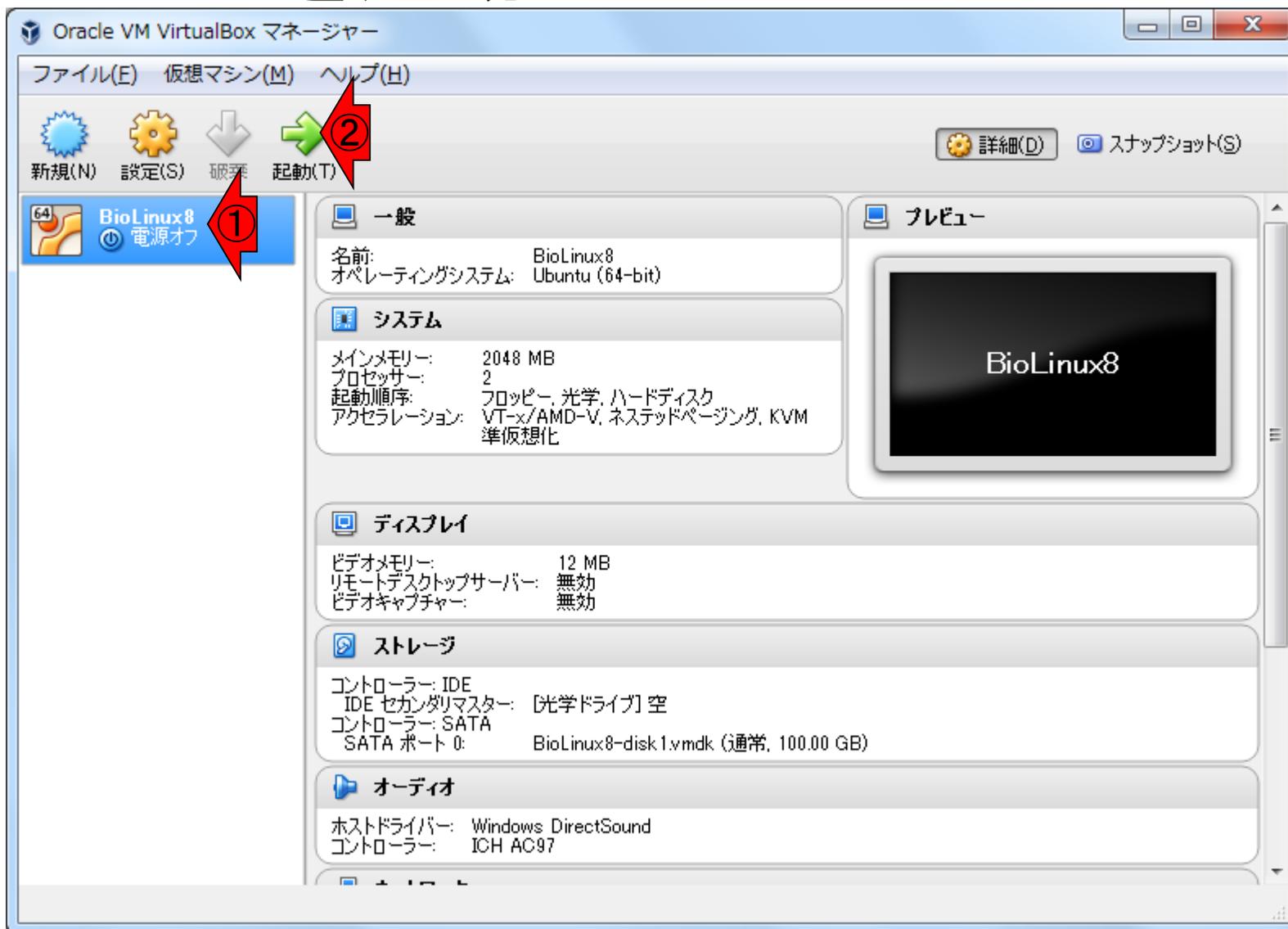


# VirtualBox起動後の状態

こんな感じになります。連載第3回ウェブ資料(JSLAB3\_suppl\_....pdf)と同じような説明

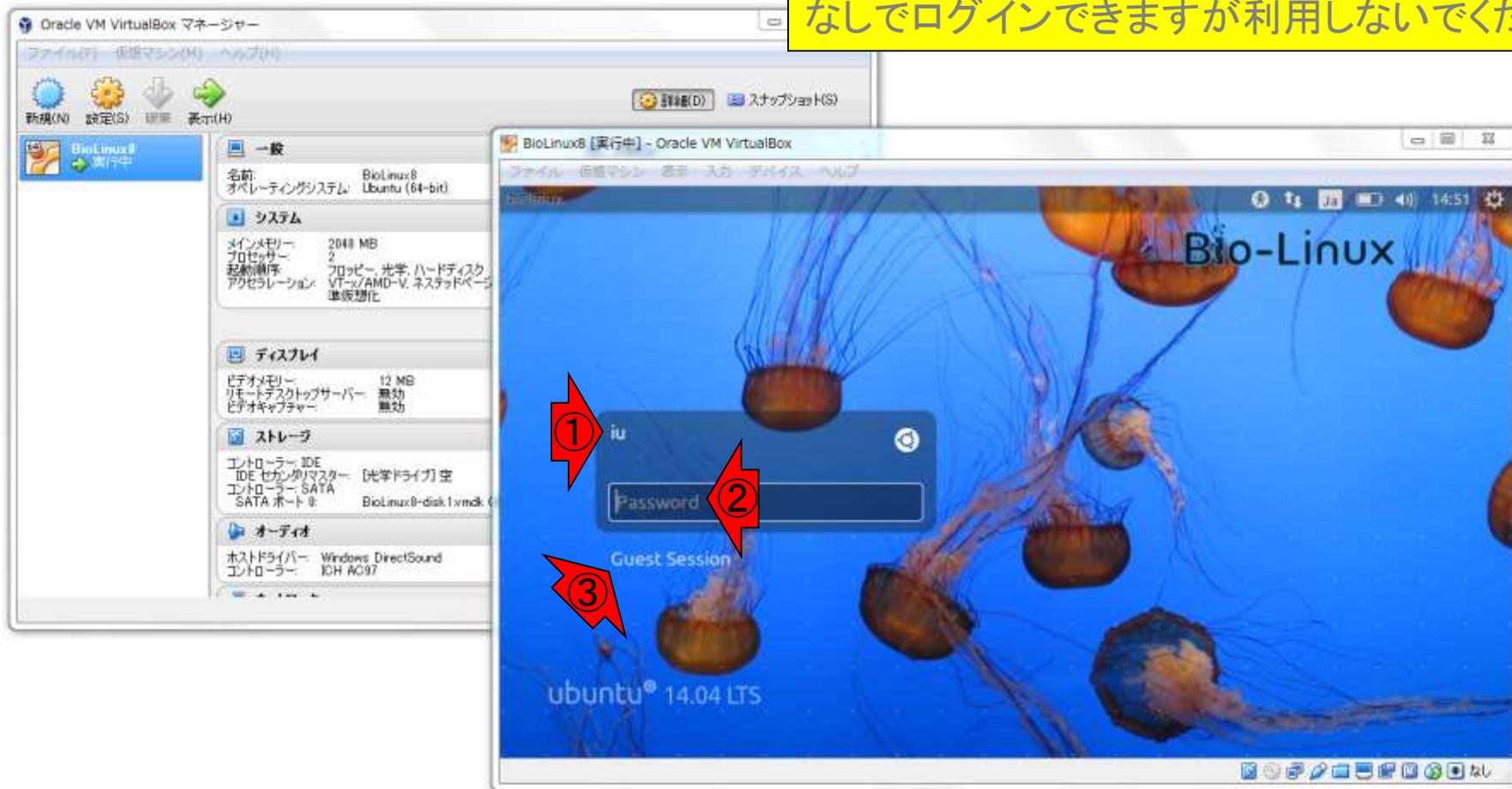


# Linuxを起動

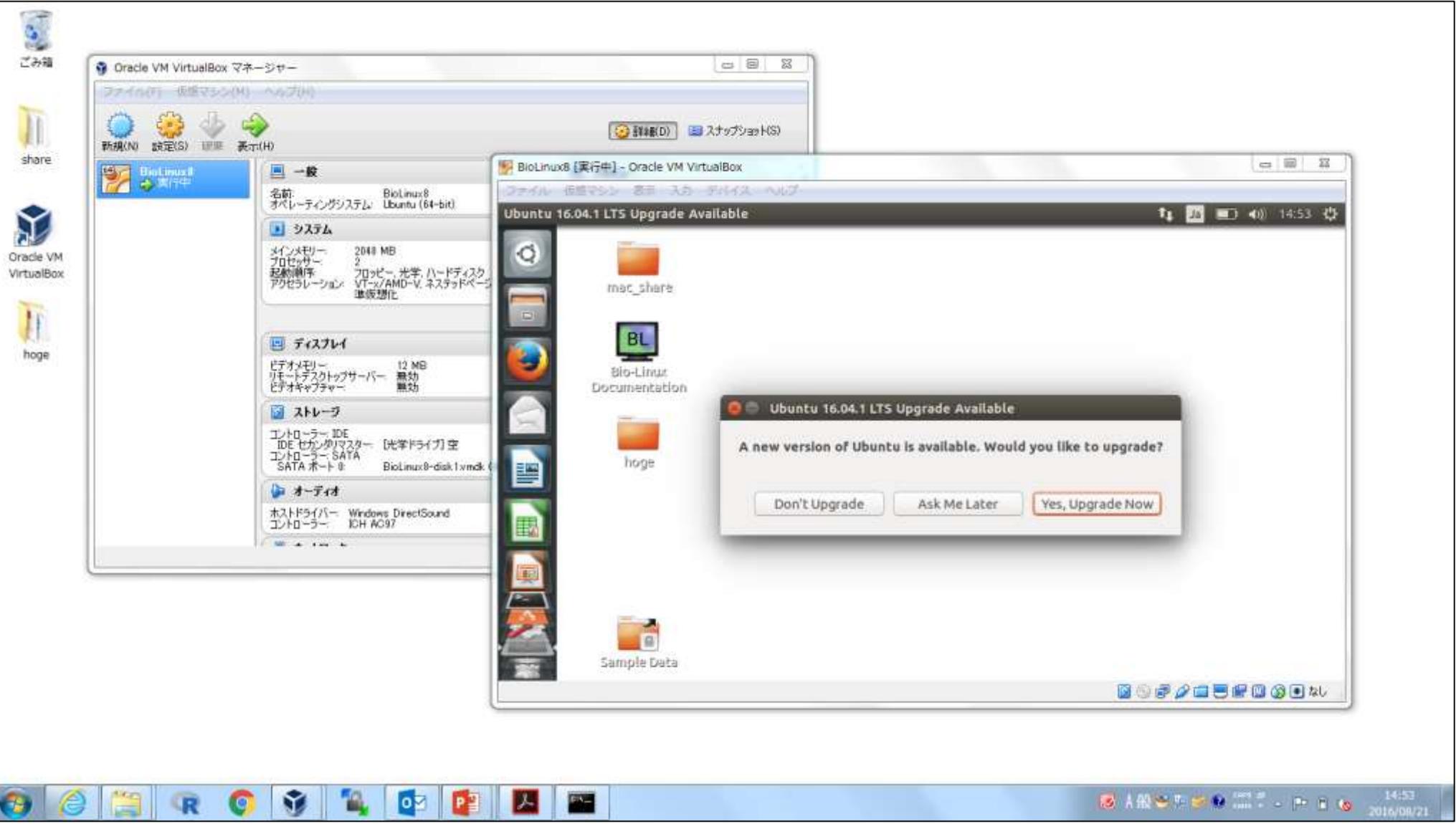


# Linux起動後の状態

約1分でこのような状態になります。Windowsのログイン画面と同じ状態です。①ユーザ名はiu、②パスワードはpass1409です。Linuxにログインしましょう。③Guest Sessionからは、パスワードなしでログインできますが利用しないでください!

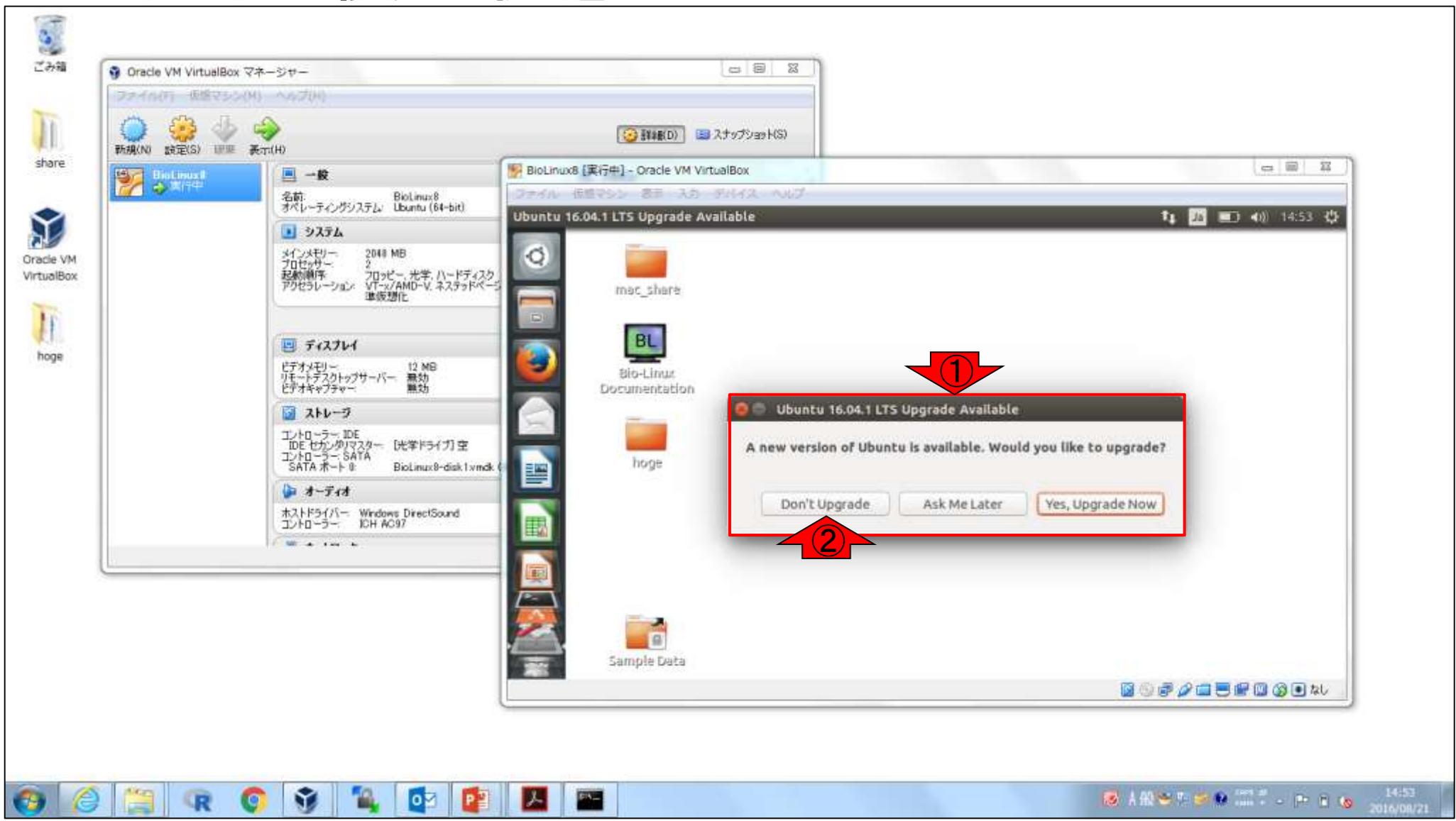


# ログイン後の状態

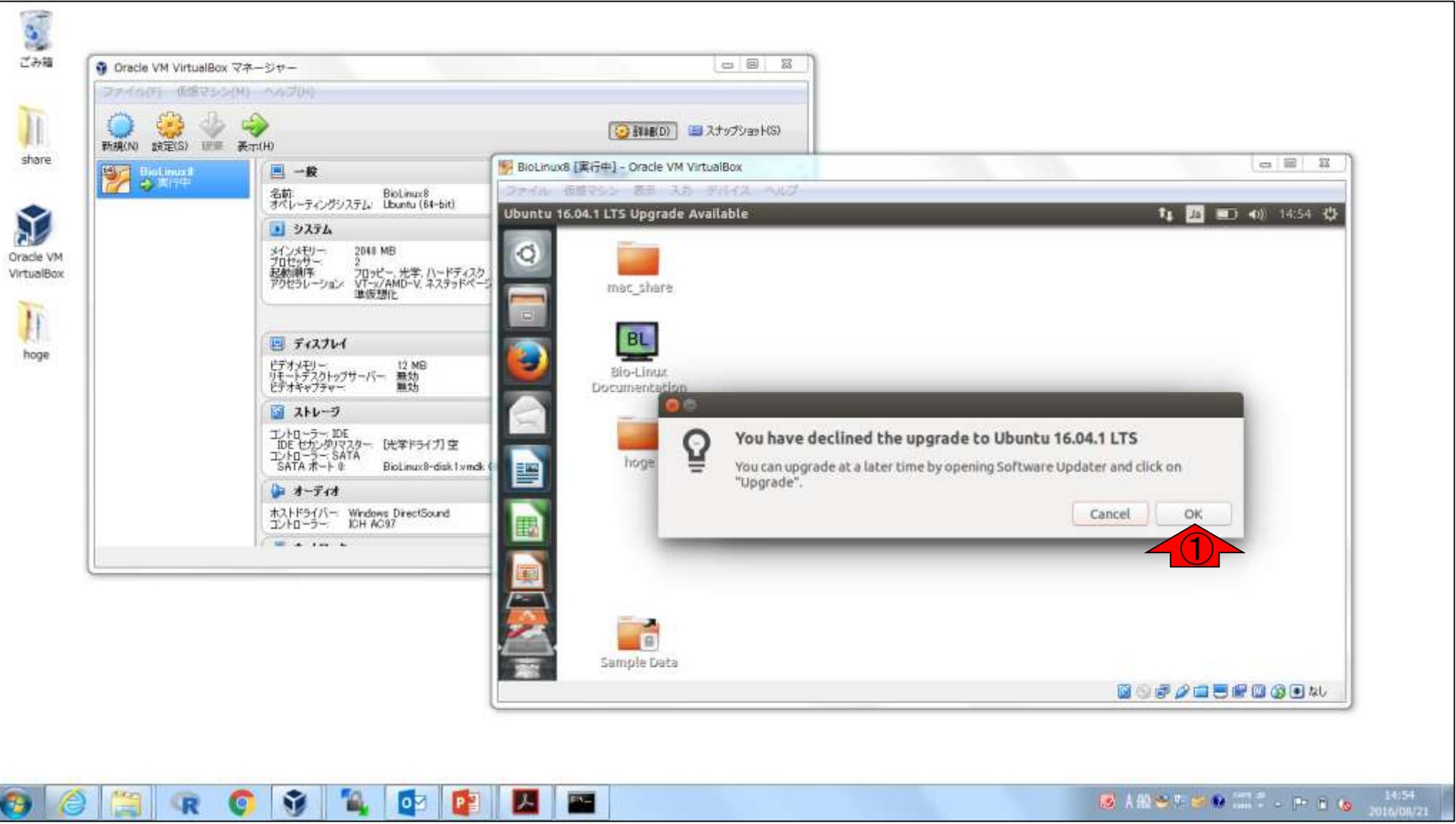


①のポップアップは「Windows10にアップグレードするか？」という類のものです。②Don't Upgrade

# ログイン後の状態

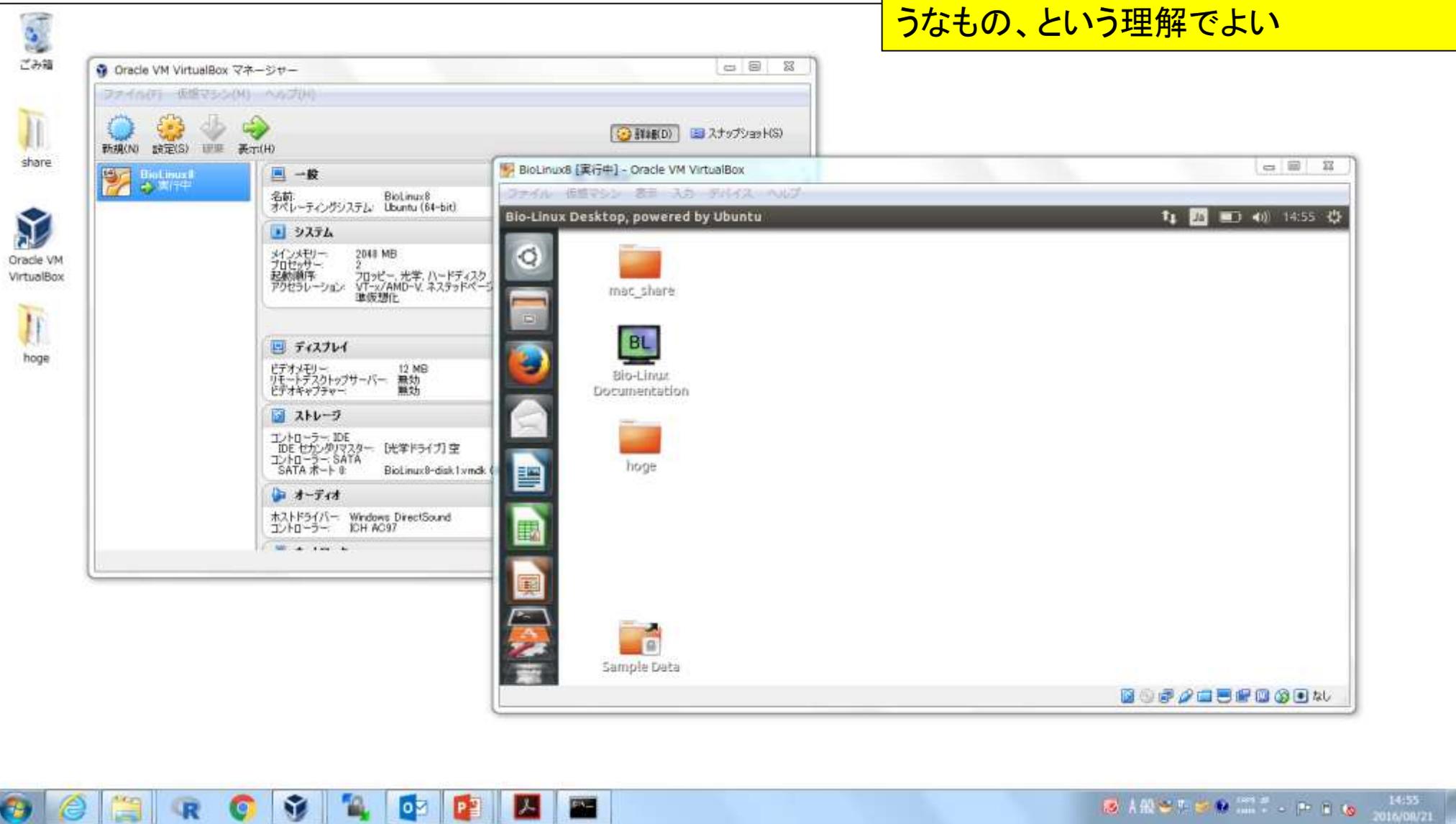


# ログイン後の状態



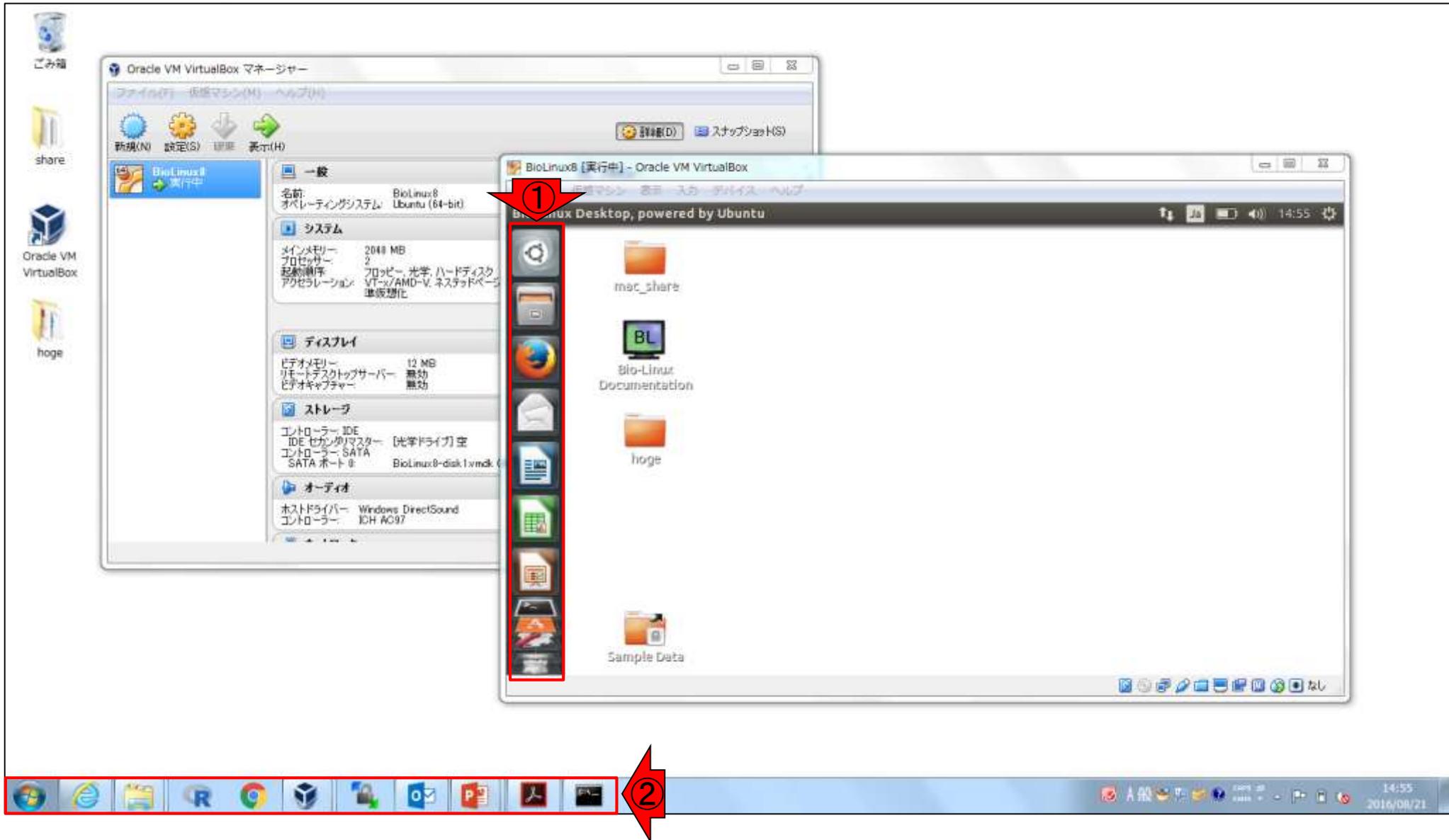
# BioLinux8起動後の状態

Windows (ホストOS) 上で、BioLinux8 というLinux (ゲストOS) が立ち上がっている状態。VirtualBoxは仲介役のようなもの、という理解でよい



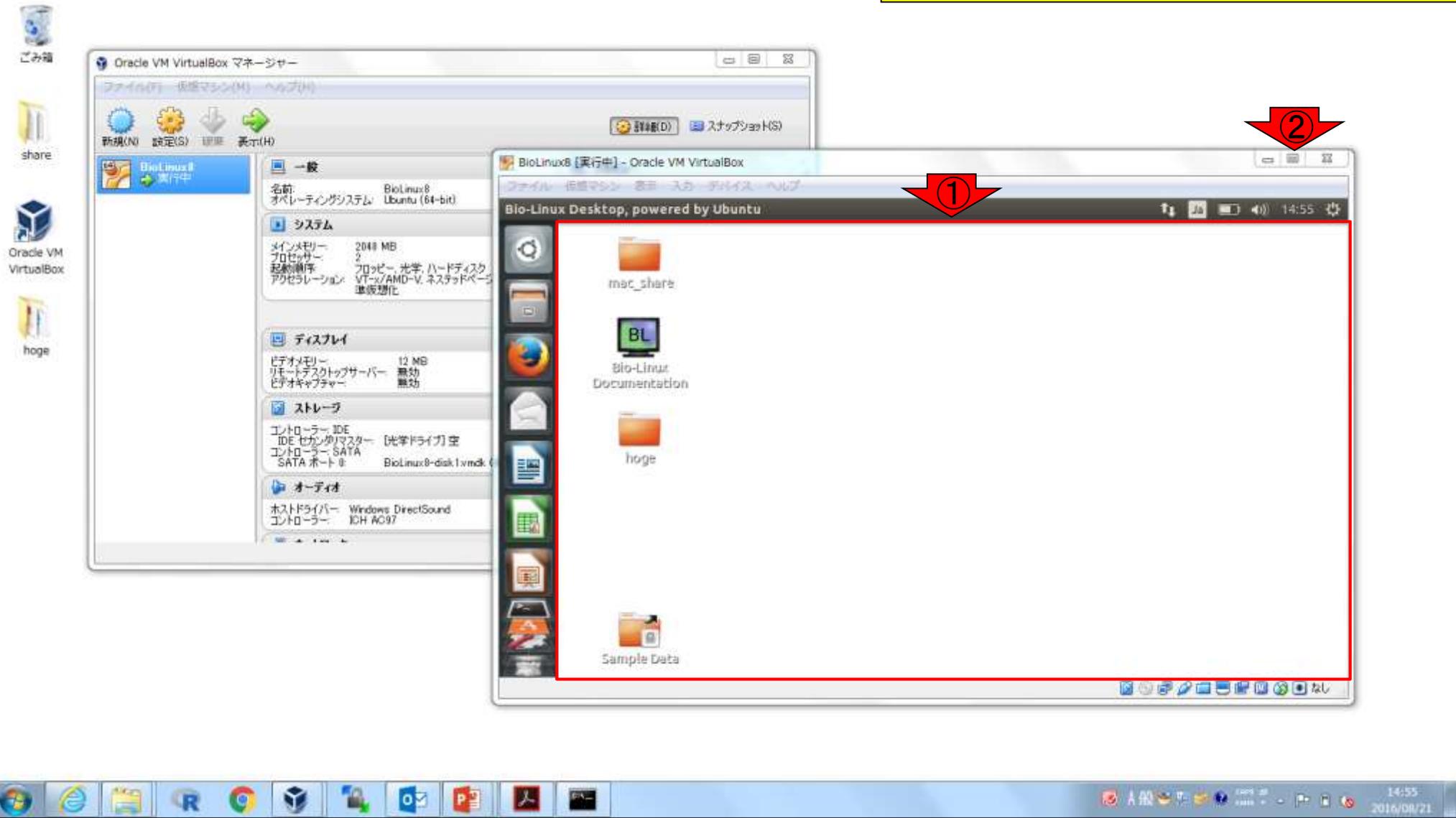
# 対応関係

①BioLinux8の赤枠部分は、②Windowsのタスクバーと同じようなもの



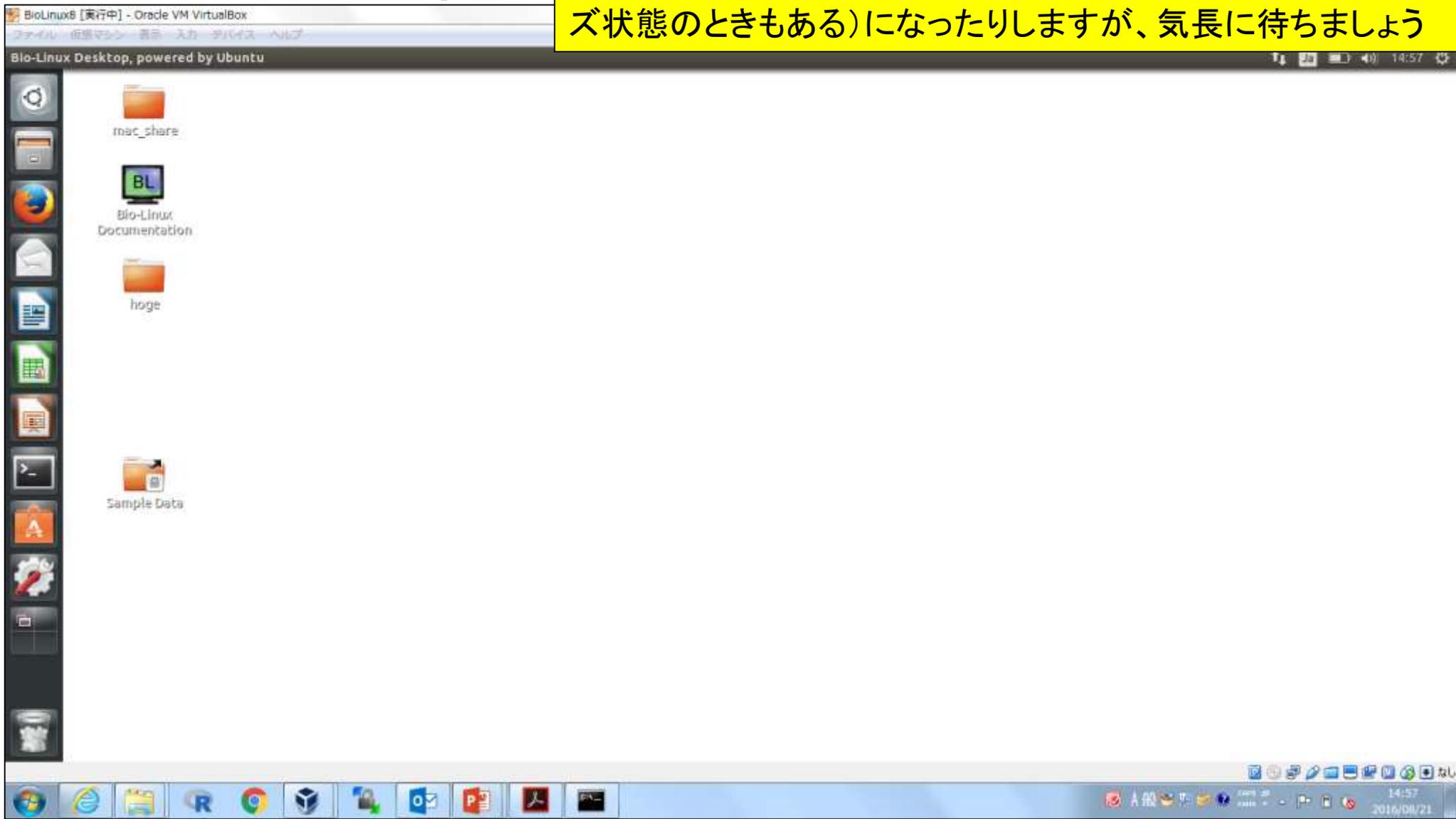
# 対応関係

①赤枠部分はBioLinux8のデスクトップ画面に相当します。②の部分を押してBioLinux8の画面を最大化すれば...

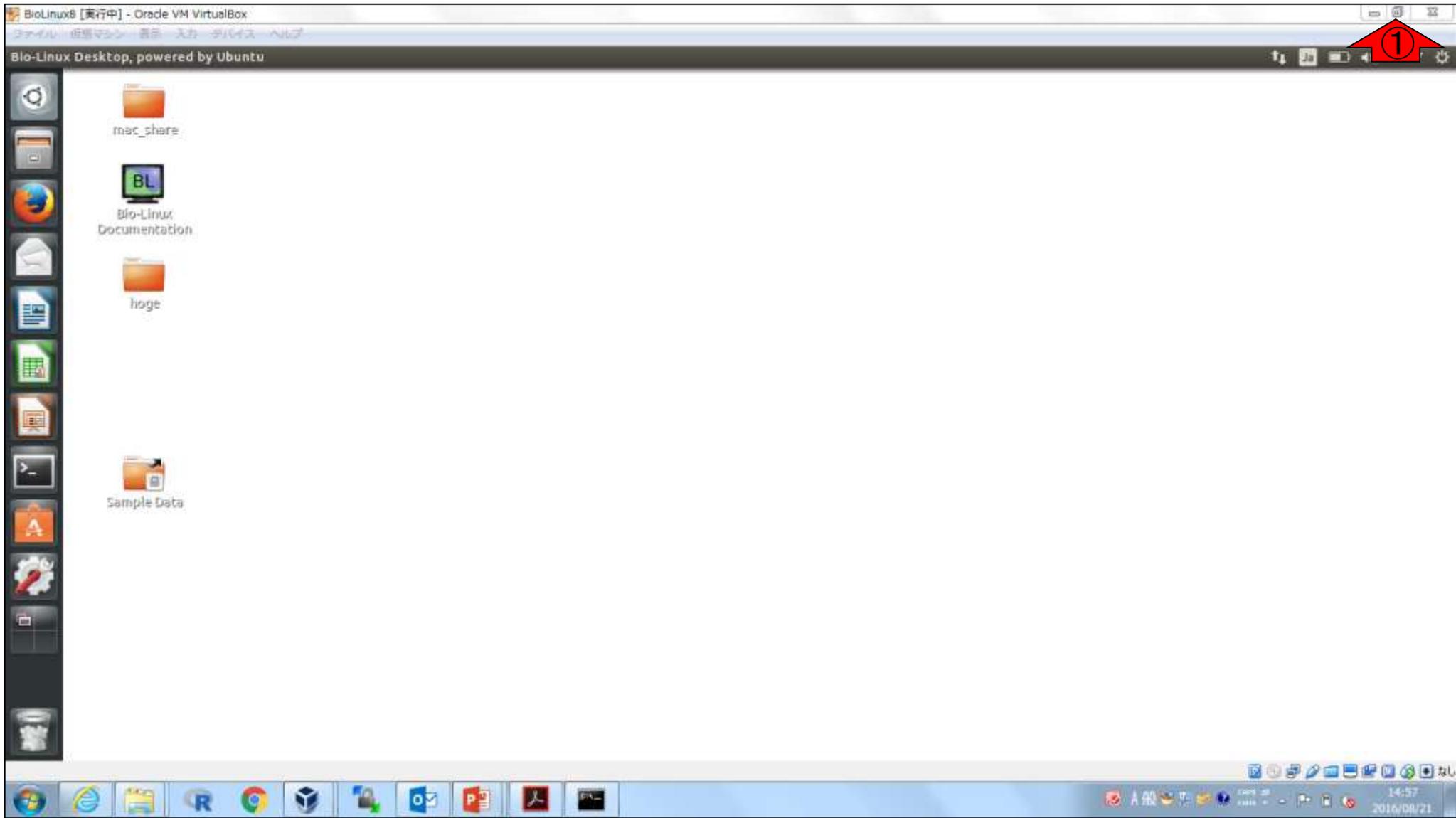


# 仮想Linux環境

仮想的にLinux環境で仕事をしているのと同じような感じになります。ただし、Windows上でLinuxを動かしているなので、どうしても動作が重くなったり、フリーズしたような感じ(本当にフリーズ状態のときもある)になりますが、気長に待ちましょう

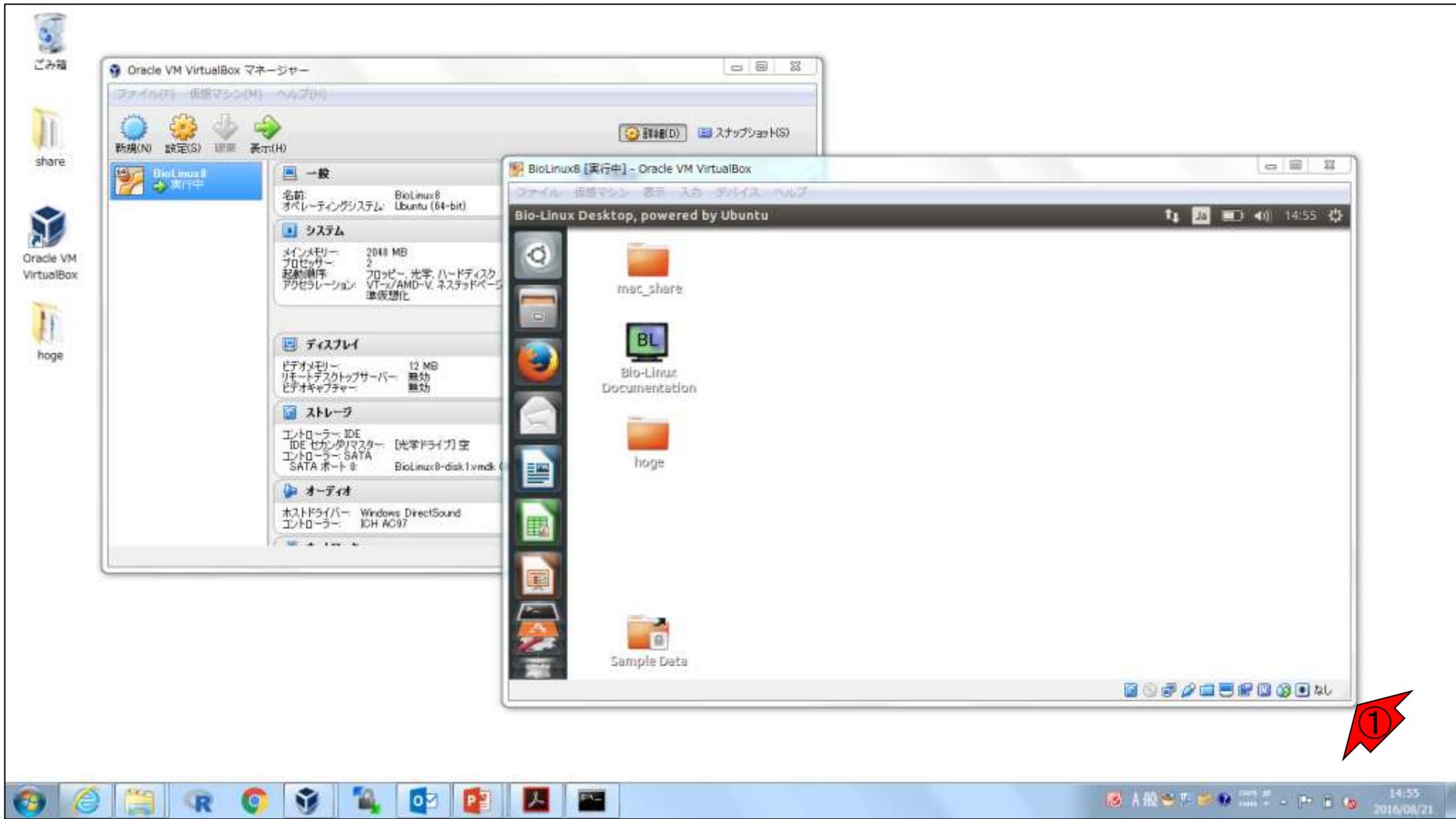


# 仮想Linux環境



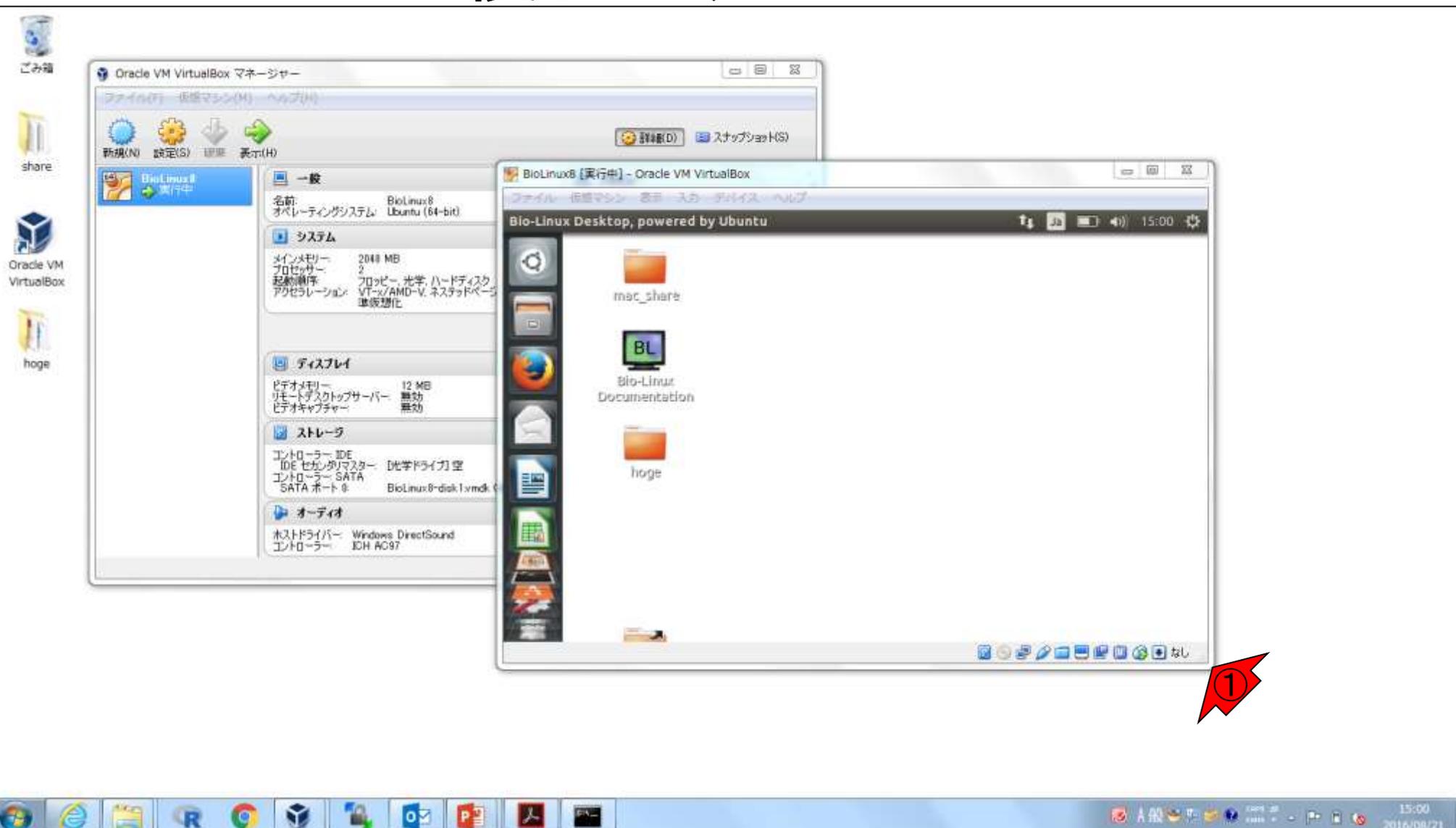
BioLinux8のGUI画面サイズを変更  
すべく、①の部分動かしてみよう

# いろいろと...慣れです



# いろいろと...慣れです

こんな感じにしたり、ガスガス変えまくっていると...動作が不安定になって落ちます



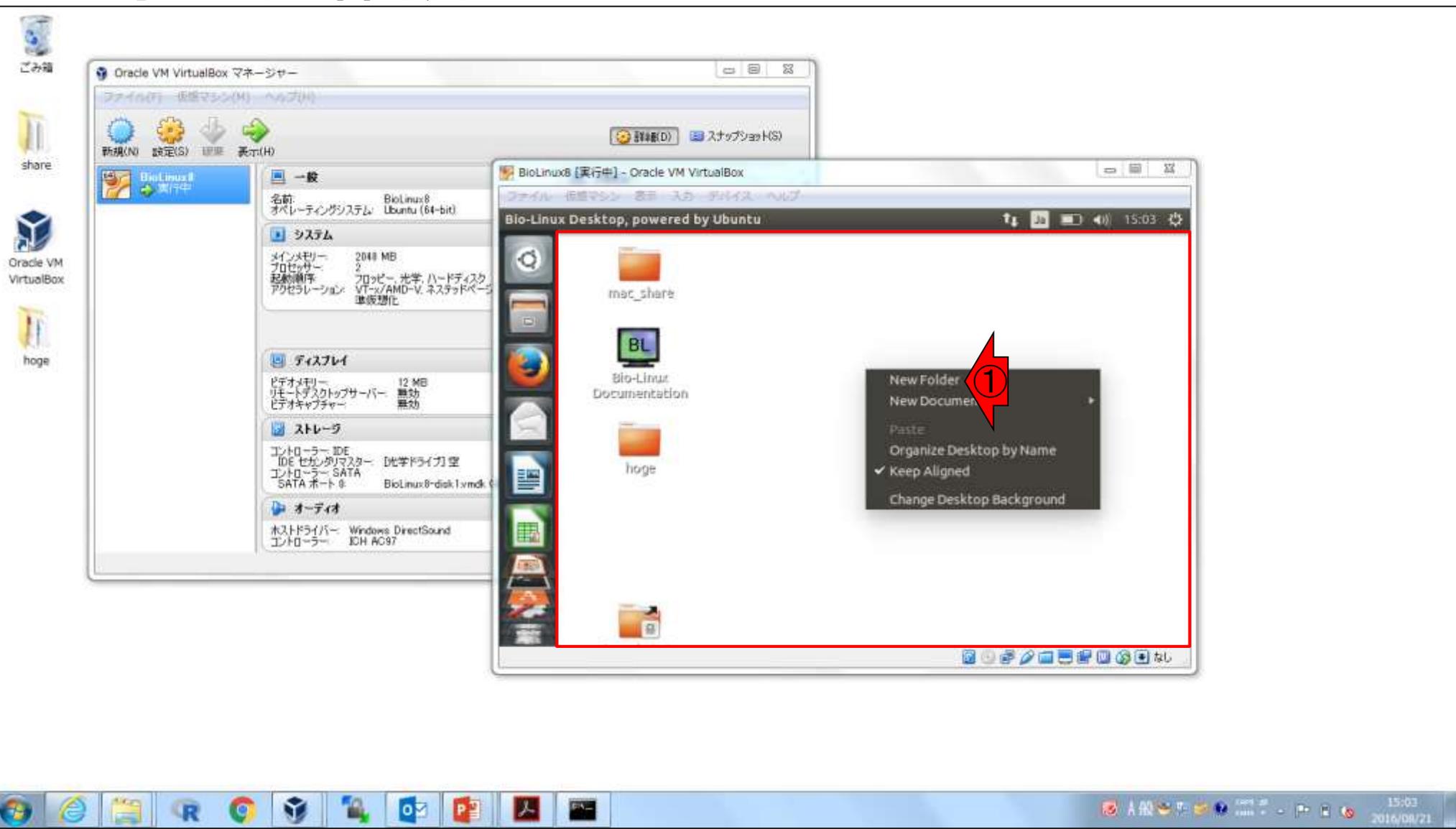
# いろいろと...慣れです

ログイン画面に戻った状態です。最初はこの程度で落ちる不安定さにイラッと思いますが、慣れです。パスワード(pass1409)を打ち込んでログインし直しましょう。この経験から、画面サイズを変更するときにはこういうことも起こるということを学んだ



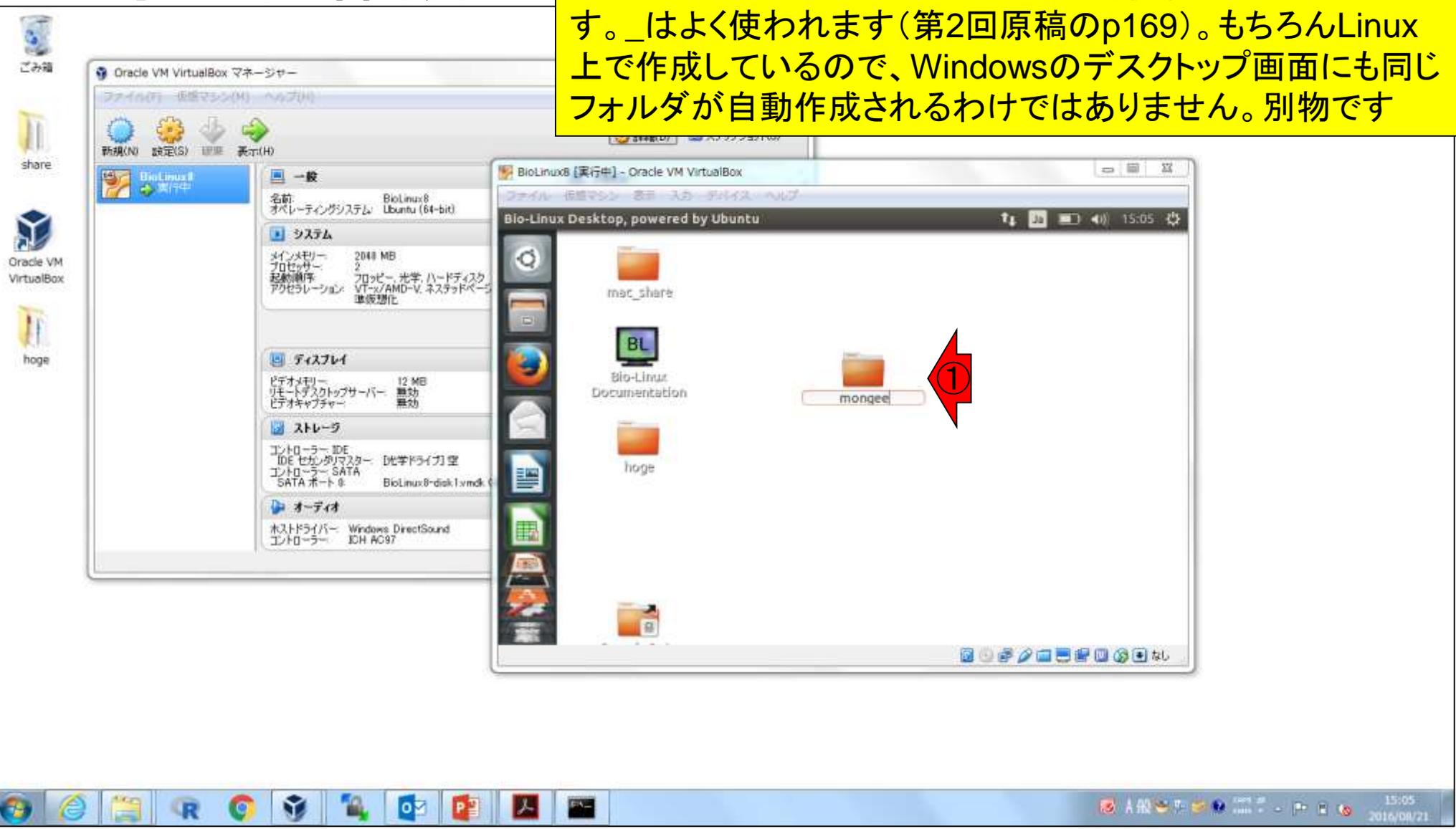
# フォルダ作成

任意の名前のフォルダを作成してみましょう。赤枠内で右クリックし、①New Folder



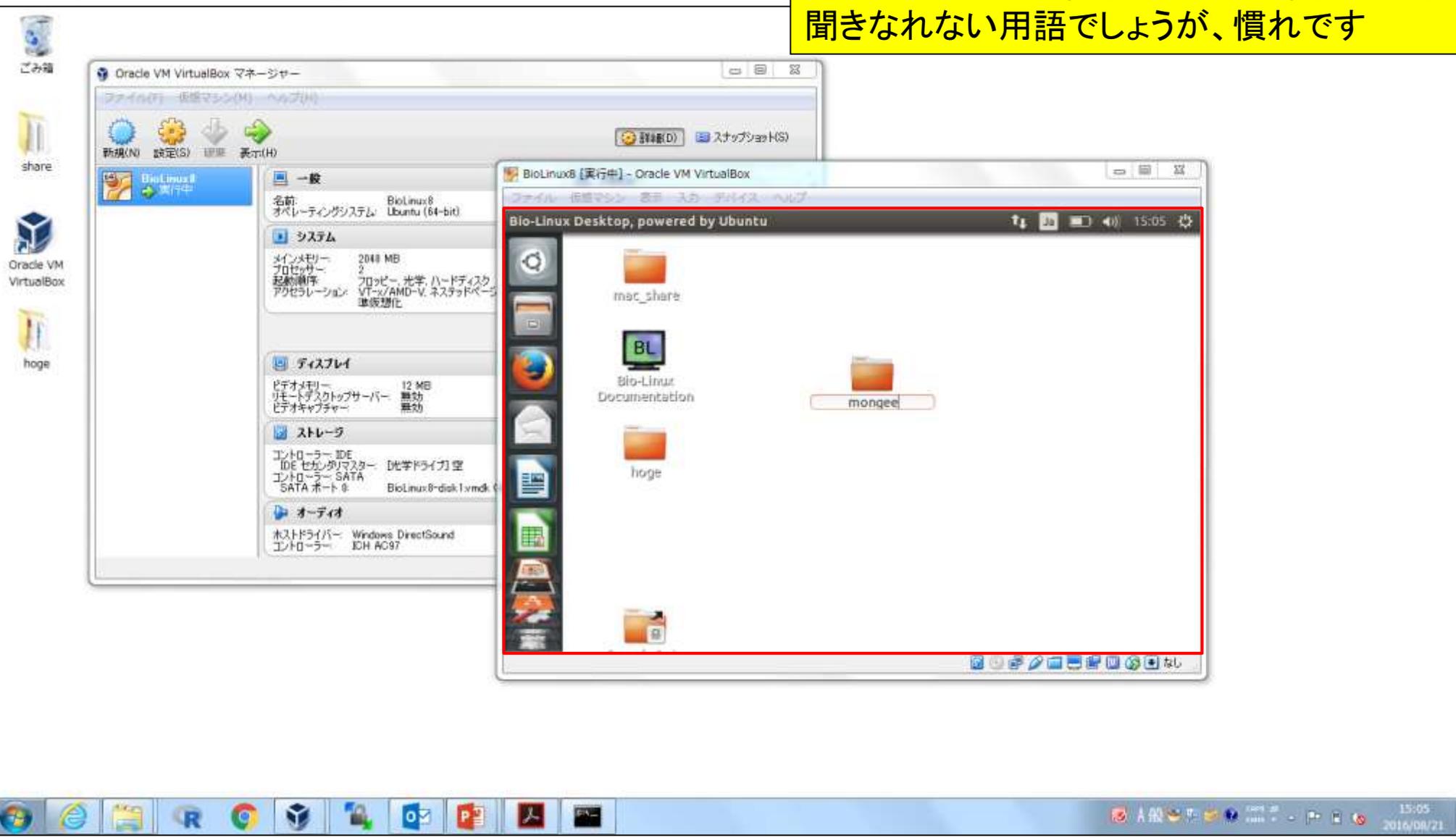
# フォルダ作成

私はmongeeというフォルダ名にしました。Linuxの世界では、フォルダ名やファイル名に、通常日本語は利用しません。また、'&%¥\*?'などの特殊文字やスペースも使わないのが常識です。\_はよく使われます(第2回原稿のp169)。もちろんLinux上で作成しているので、Windowsのデスクトップ画面にも同じフォルダが自動作成されるわけではありません。別物です



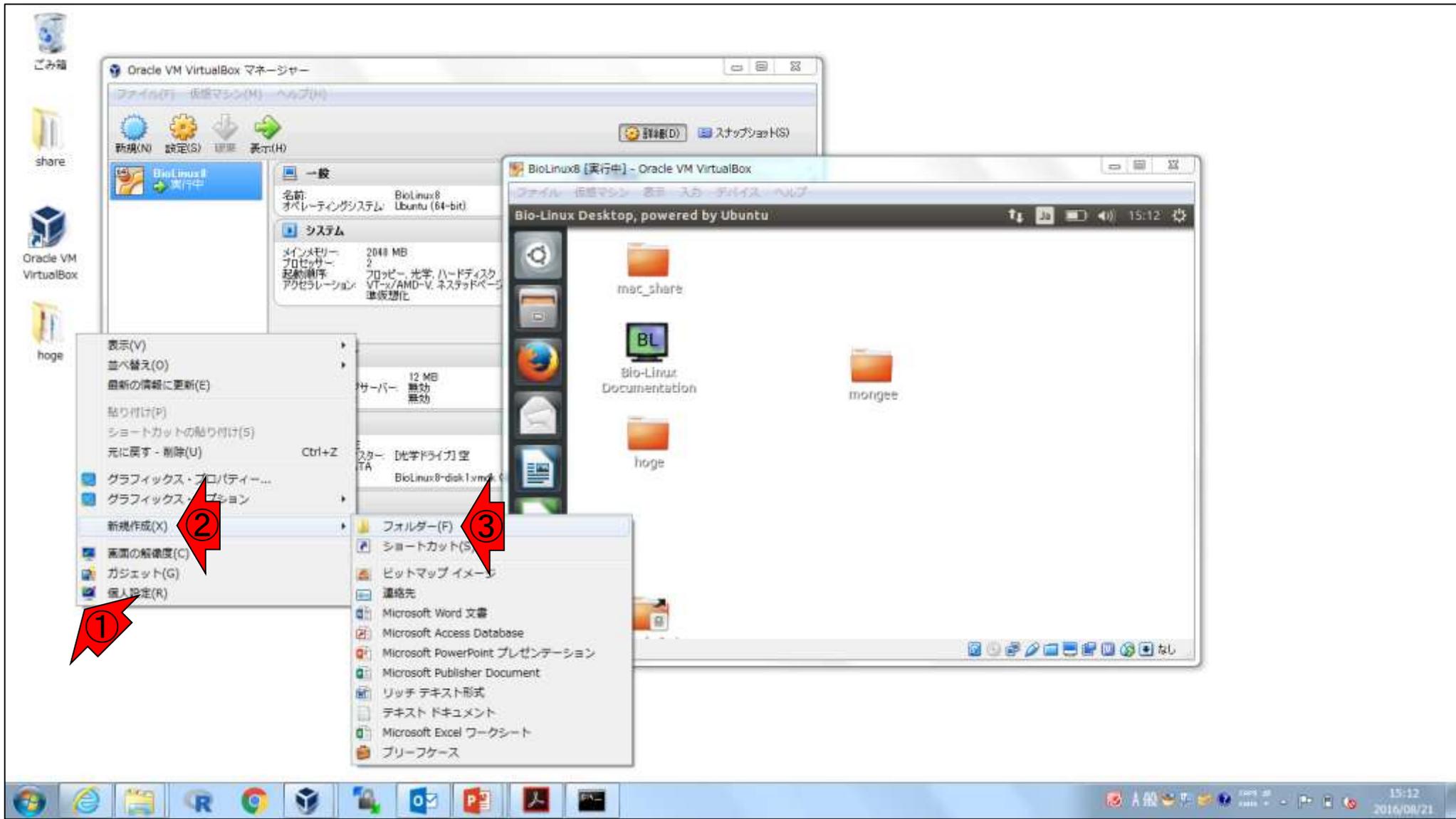
# ゲストとホスト

今はWindows上でLinuxを動かしています。  
赤枠内がLinux環境(ゲストOS環境)で、それ  
以外がWindows環境(ホストOS環境)です。  
聞きなれない用語でしょうが、慣れです



Windows(ホストOS環境)上で、「①右クリック、②新規作成、③フォルダー」の流れで新規フォルダの作成が可能です

# 念のため



# Contents

## ■ イン트로ダクション

- 概要、背景(NGS用カリキュラム、講習会)、Linuxスキル習得の意義
- ウェブ情報(日本乳酸菌学会誌のNGS連載やNGS講習会資料)

## ■ 実習環境に慣れる

- 仮想環境での作業に慣れる
- GUIとCUI(マウス操作かコマンド入力操作か)
- ターミナルでの作業
- 共有フォルダの概念を理解

## ■ 練習

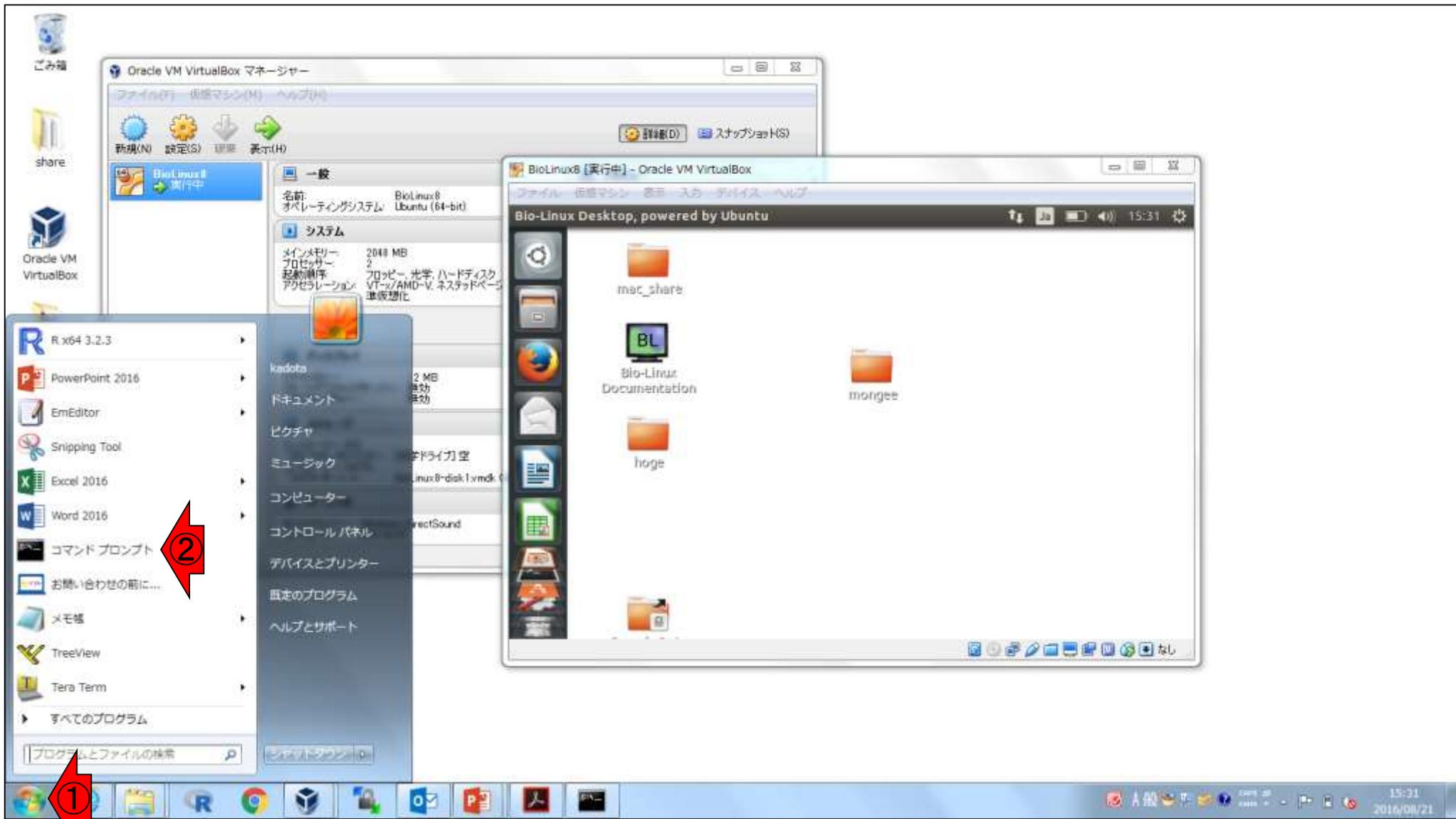
- 作業ディレクトリの変更、練習用NGSデータファイルのダウンロード
- ファイルの確認、de novoゲノムアセンブリ
- BLAST検索

## ■ 課題

- グループごとに異なる課題ファイルを入力として、「ダウンロード、de novoアセンブリ、BLAST検索」を実行し、得られた結果をレポートにまとめて発表せよ
- グループ1はkadai1.fasta、グループ2はkadai2.fasta、etc.

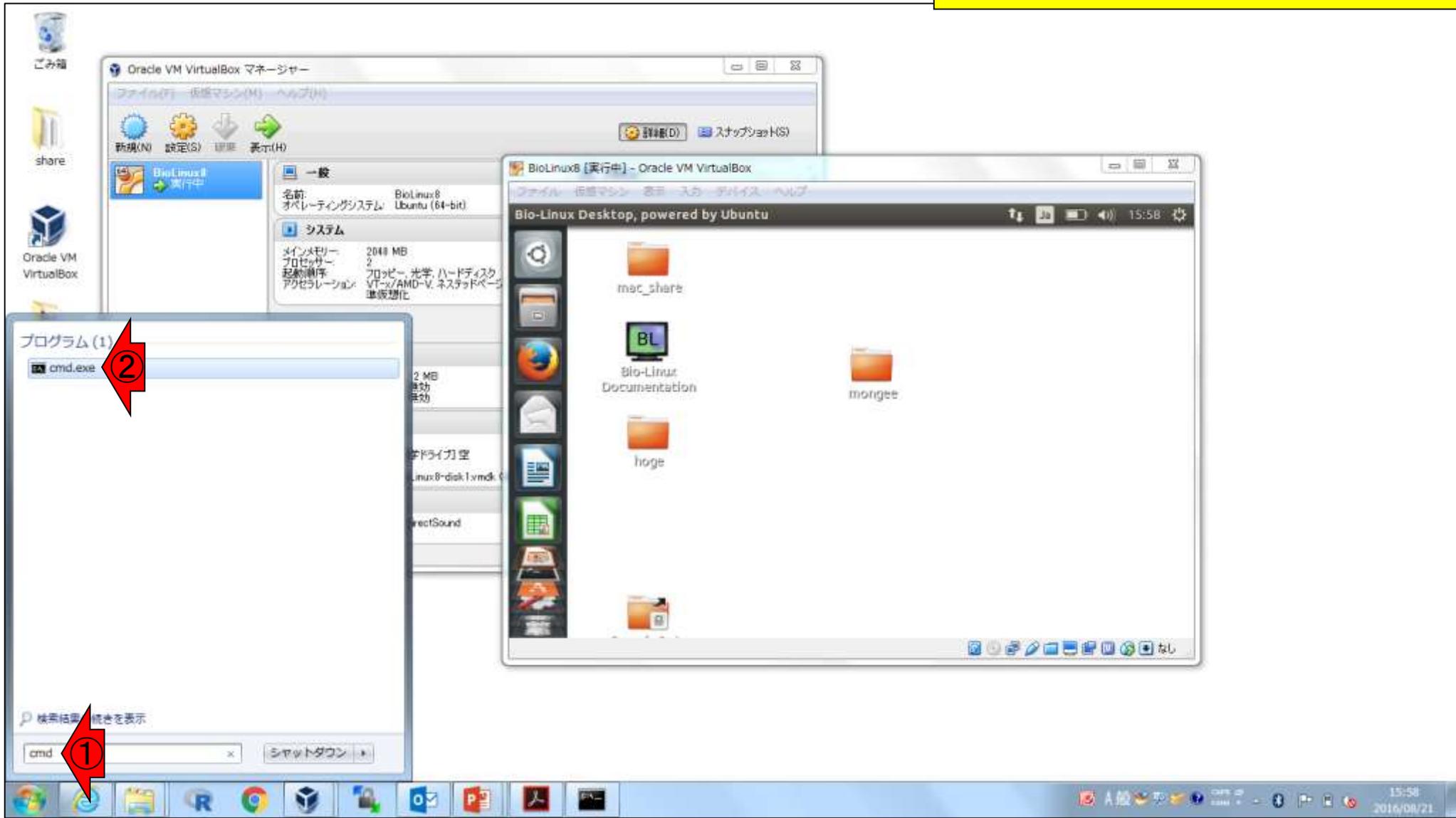
# GUIとCUI

①スタートメニューから、②コマンドプロンプトを選んで起動しましょう



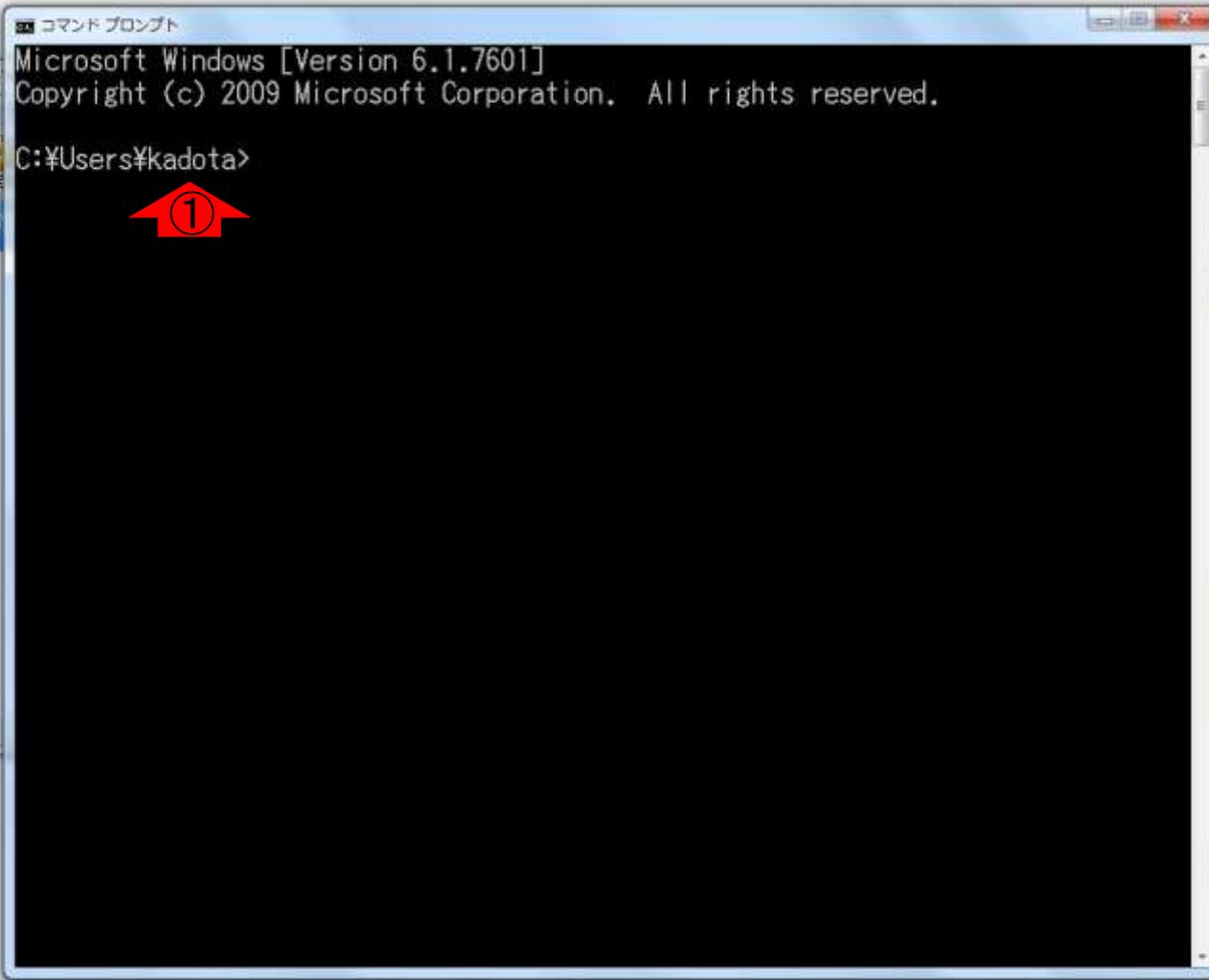
# GUIとCUI

コマンドプロンプトがすぐに見つからない場合は、①検索窓で、cmdと打つのもよいです。②cmd.exe



# GUIとCUI

コマンドプロンプト起動後の状態。貸与PCはユーザ名iuなので、①の部分が「C:\Users\iu」。Macintoshのヒトは、「ターミナル」を起動するのと同じと思えばよい

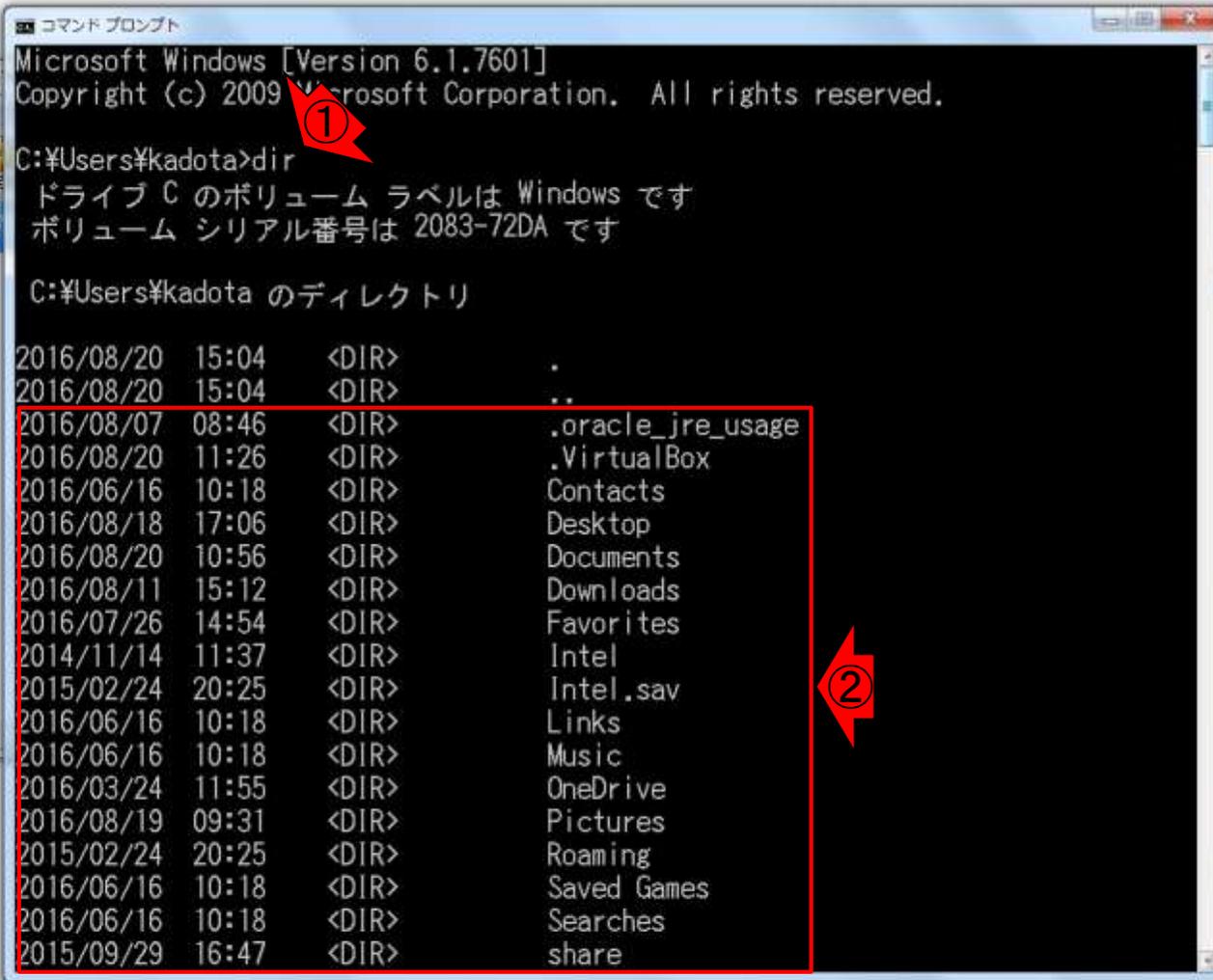


```
コマンド プロンプト
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\kadota>
```

# GUIとCUI

- ①dirと打って、リターンキーを押す。
- ②赤枠で見ているものは...



```
コマンド プロンプト
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

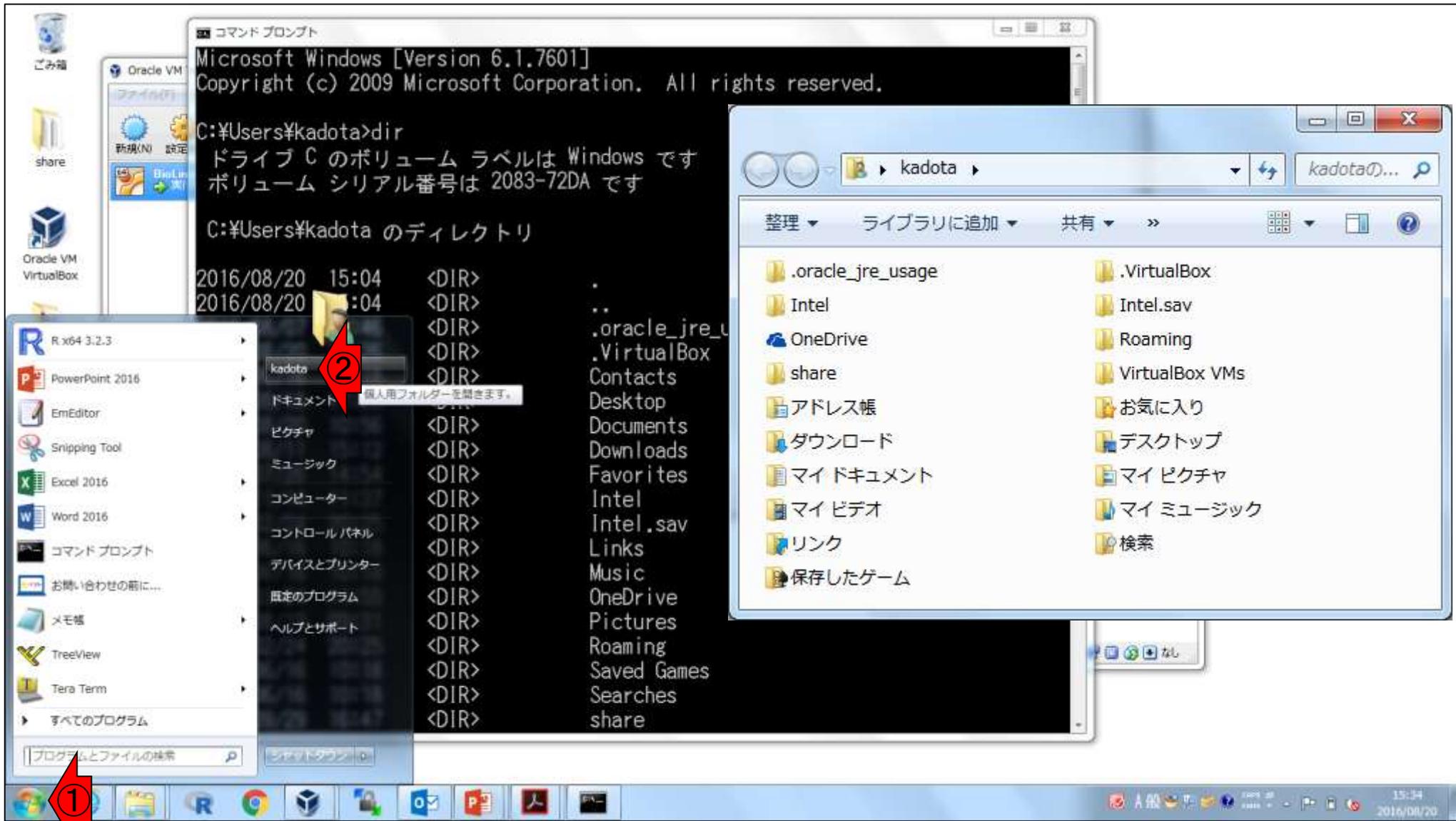
C:\Users\kadota>dir
ドライブ C のボリューム ラベルは Windows です
ボリューム シリアル番号は 2083-72DA です

C:\Users\kadota のディレクトリ

2016/08/20  15:04    <DIR>          .
2016/08/20  15:04    <DIR>          ..
2016/08/07  08:46    <DIR>          .oracle_jre_usage
2016/08/20  11:26    <DIR>          .VirtualBox
2016/06/16  10:18    <DIR>          Contacts
2016/08/18  17:06    <DIR>          Desktop
2016/08/20  10:56    <DIR>          Documents
2016/08/11  15:12    <DIR>          Downloads
2016/07/26  14:54    <DIR>          Favorites
2014/11/14  11:37    <DIR>          Intel
2015/02/24  20:25    <DIR>          Intel.sav
2016/06/16  10:18    <DIR>          Links
2016/06/16  10:18    <DIR>          Music
2016/03/24  11:55    <DIR>          OneDrive
2016/08/19  09:31    <DIR>          Pictures
2015/02/24  20:25    <DIR>          Roaming
2016/06/16  10:18    <DIR>          Saved Games
2016/06/16  10:18    <DIR>          Searches
2015/09/29  16:47    <DIR>          share
```

# GUIとCUI

①スタートメニューの右上にある、②ユーザ名 kadotaの「ホームディレクトリ」の中身です



# GUIとCUI

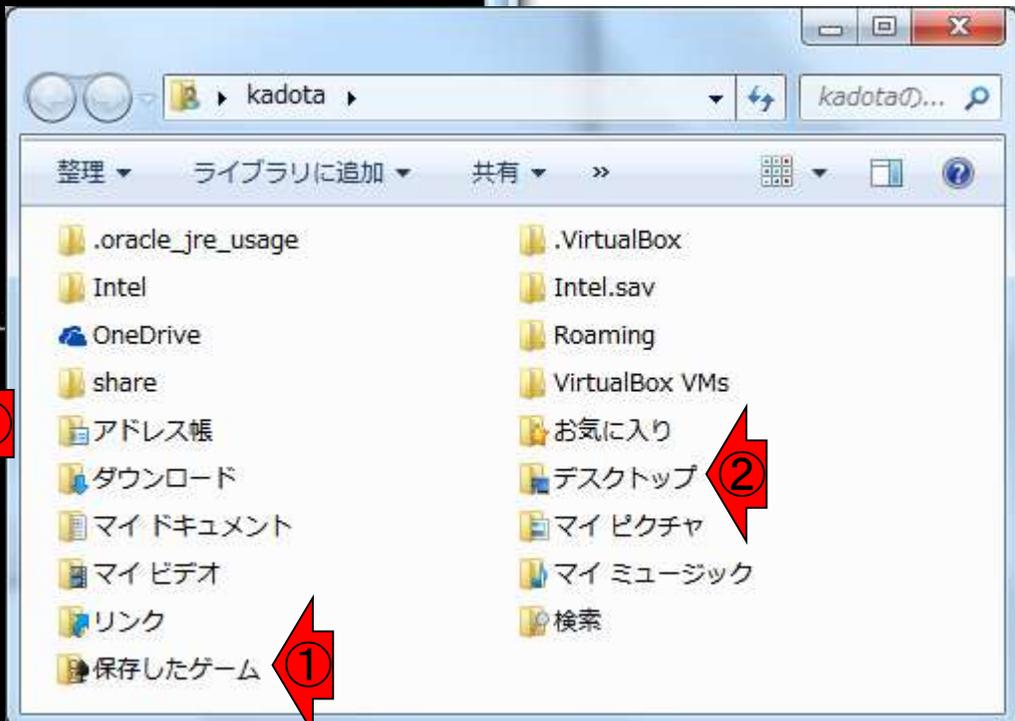
①「Saved Games ⇔ 保存したゲーム」、②「Desktop ⇔ デスクトップ」などと「English ⇔ 日本語」の変換が自動でなされていますが、これはWindows側でよきに計らってくれているためと思えばよいです

```
コマンド プロンプト
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:¥Users¥kadota>dir
ドライブ C のボリューム ラベルは Windows です
ボリューム シリアル番号は 2083-72DA です

C:¥Users¥kadota のディレクトリ

2016/08/20  15:04  <DIR>          .
2016/08/20  15:04  <DIR>          ..
2016/08/07  08:46  <DIR>          .oracle_jre_u
2016/08/20  11:26  <DIR>          .VirtualBox
2016/06/16  10:18  <DIR>          Contacts
2016/08/18  17:06  <DIR>          Desktop
2016/08/20  10:56  <DIR>          Documents
2016/08/11  15:12  <DIR>          Downloads
2016/07/26  14:54  <DIR>          Favorites
2014/11/14  11:37  <DIR>          Intel
2015/02/24  20:25  <DIR>          Intel.sav
2016/06/16  10:18  <DIR>          Links
2016/06/16  10:18  <DIR>          Music
2016/03/24  11:55  <DIR>          OneDrive
2016/08/19  09:31  <DIR>          Pictures
2015/02/24  20:25  <DIR>          Roaming
2016/06/16  10:18  <DIR>          Saved Games
2016/06/16  10:18  <DIR>          Searches
2015/09/29  16:47  <DIR>          share
```



②

②

①

①

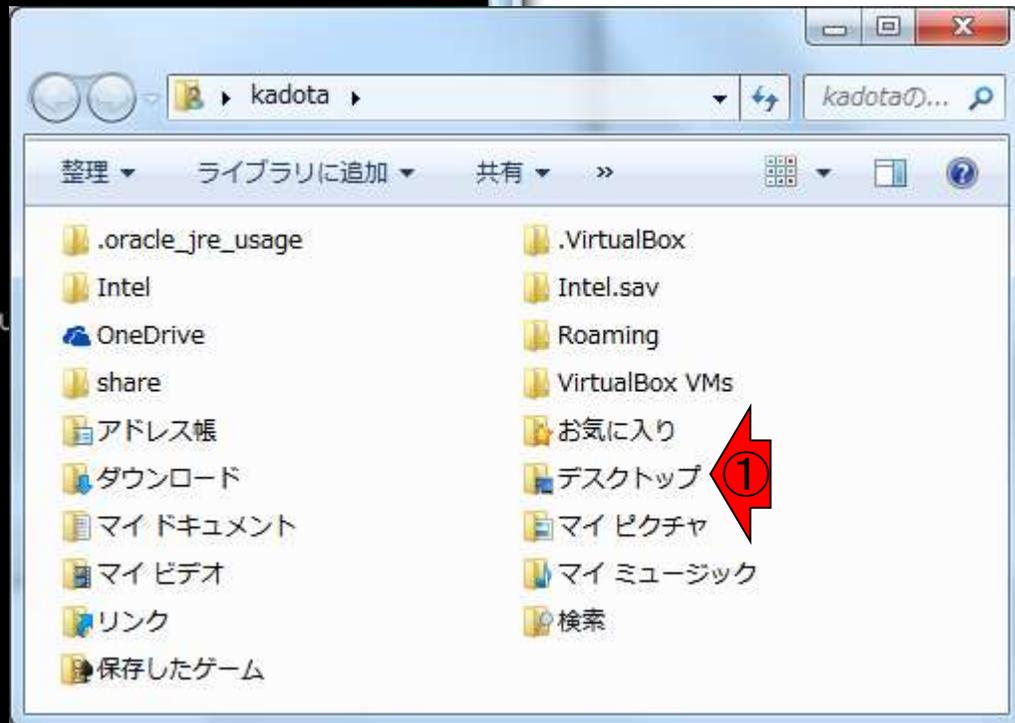
# GUIとCUI

```
コマンド プロンプト
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\kadota>dir
ドライブ C のボリューム ラベルは Windows です
ボリューム シリアル番号は 2083-72DA です

C:\Users\kadota のディレクトリ

2016/08/20  15:04  <DIR>      .
2016/08/20  15:04  <DIR>      ..
2016/08/07  08:46  <DIR>      .oracle_jre_u
2016/08/20  11:26  <DIR>      .VirtualBox
2016/06/16  10:18  <DIR>      Contacts
2016/08/18  17:06  <DIR>      Desktop
2016/08/20  10:56  <DIR>      Documents
2016/08/11  15:12  <DIR>      Downloads
2016/07/26  14:54  <DIR>      Favorites
2014/11/14  11:37  <DIR>      Intel
2015/02/24  20:25  <DIR>      Intel.sav
2016/06/16  10:18  <DIR>      Links
2016/06/16  10:18  <DIR>      Music
2016/03/24  11:55  <DIR>      OneDrive
2016/08/19  09:31  <DIR>      Pictures
2015/02/24  20:25  <DIR>      Roaming
2016/06/16  10:18  <DIR>      Saved Games
2016/06/16  10:18  <DIR>      Searches
2015/09/29  16:47  <DIR>      share
```



# GUIとCUI

①kadotaのPC環境では、②赤枠の3つしかない  
ので、それに相当するものが③で見えています

コマンド プロンプト

```
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\kadota>dir
ドライブ C のボリューム ラベルは Windows です
ボリューム シリアル番号は 2083-72DA です

C:\Users\kadota のディレクトリ

2016/08/20  15:04  <DIR>          .
2016/08/20  15:04  <DIR>          ..
2016/08/07  08:46  <DIR>          .oracle_jre_u
2016/08/20  11:26  <DIR>          .VirtualBox
2016/06/16  10:18  <DIR>          Contacts
2016/08/18  17:06  <DIR>          Desktop
2016/08/20  10:56  <DIR>          Documents
2016/08/11  15:12  <DIR>          Downloads
2016/07/26  14:54  <DIR>          Favorites
2014/11/14  11:37  <DIR>          Intel
2015/02/24  20:25  <DIR>          Intel.sav
2016/06/16  10:18  <DIR>          Links
2016/06/16  10:18  <DIR>          Music
2016/03/24  11:55  <DIR>          OneDrive
2016/08/19  09:31  <DIR>          Pictures
2015/02/24  20:25  <DIR>          Roaming
2016/06/16  10:18  <DIR>          Saved Games
2016/06/16  10:18  <DIR>          Searches
2015/09/29  16:47  <DIR>          share
```

File Explorer: C:\Users\kadota\Desktop

名前	更新日時	種類
hoge	2016/08/18 11:36	ファイル フォル...
share	2016/07/11 10:45	ファイル フォル...
Oracle VM VirtualBox	2015/11/19 17:30	ショートカット

# GUIとCUI

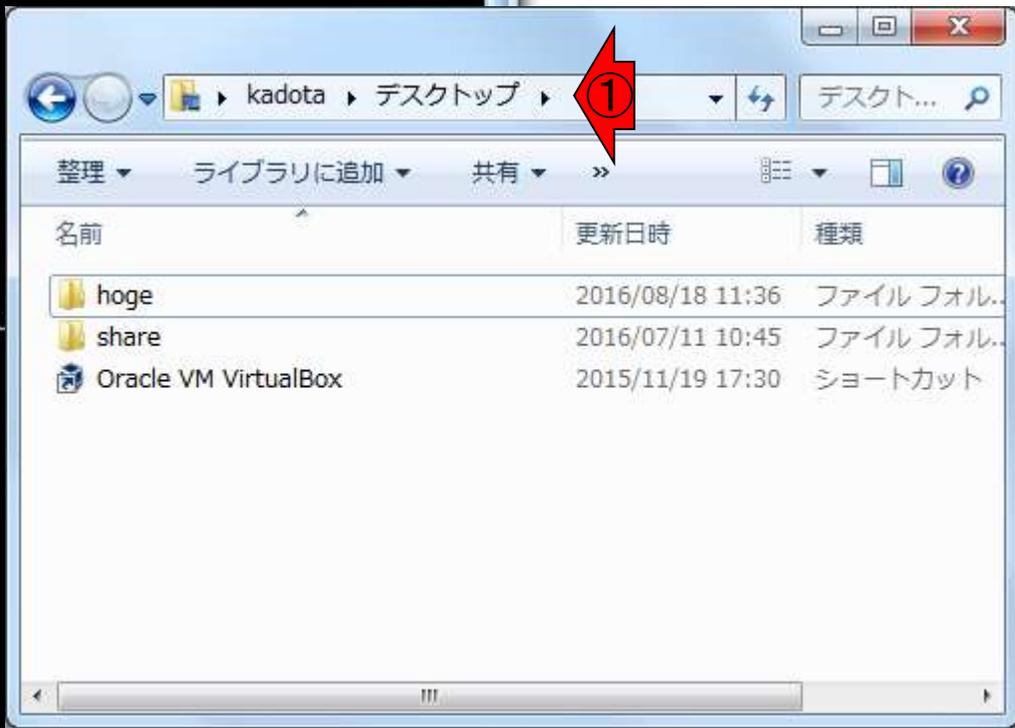
「kadotaさんのデスクトップ」であることが①で明示されているので、現在どこで作業をしているかがよくわかります。①の場所を「作業ディレクトリ (working directory)」や「カレントディレクトリ (current directory)」などと呼びます。フォルダとディレクトリは、同じようなものという理解でよい

```
コマンド プロンプト
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation

C:\Users\kadota>dir
ドライブ C のボリューム ラベルは Windows です
ボリューム シリアル番号は 2083-72DA です

C:\Users\kadota のディレクトリ

2016/08/20  15:04  <DIR>          .
2016/08/20  15:04  <DIR>          ..
2016/08/07  08:46  <DIR>          .oracle_jre_u
2016/08/20  11:26  <DIR>          .VirtualBox
2016/06/16  10:18  <DIR>          Contacts
2016/08/18  17:06  <DIR>          Desktop
2016/08/20  10:56  <DIR>          Documents
2016/08/11  15:12  <DIR>          Downloads
2016/07/26  14:54  <DIR>          Favorites
2014/11/14  11:37  <DIR>          Intel
2015/02/24  20:25  <DIR>          Intel.sav
2016/06/16  10:18  <DIR>          Links
2016/06/16  10:18  <DIR>          Music
2016/03/24  11:55  <DIR>          OneDrive
2016/08/19  09:31  <DIR>          Pictures
2015/02/24  20:25  <DIR>          Roaming
2016/06/16  10:18  <DIR>          Saved Games
2016/06/16  10:18  <DIR>          Searches
2015/09/29  16:47  <DIR>          share
```



# GUIとCUI

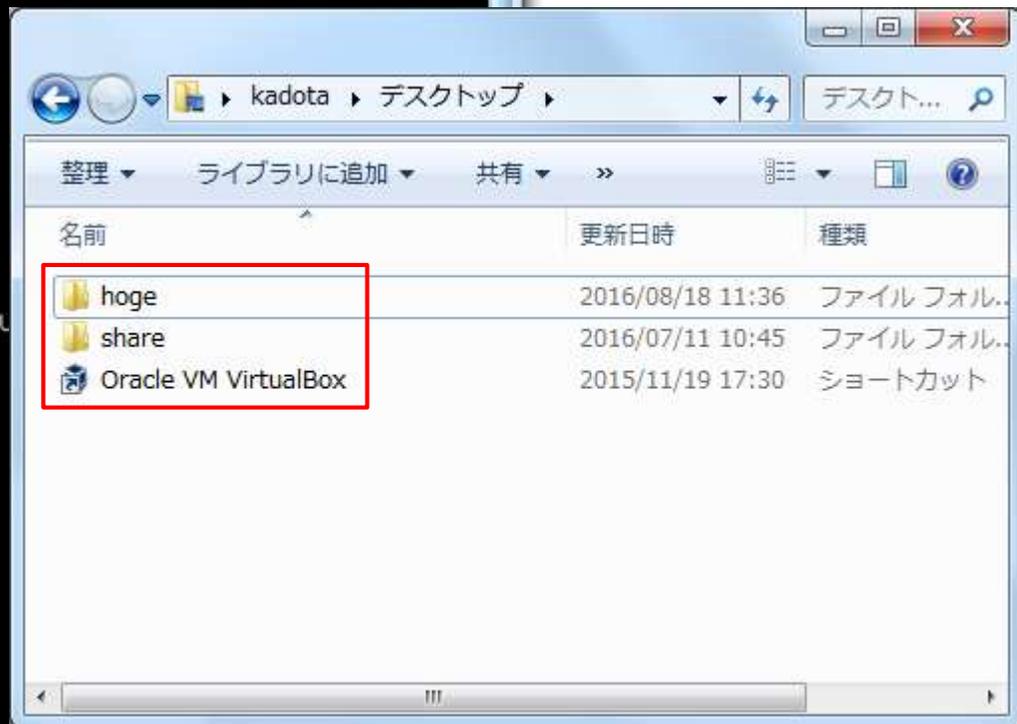
GUI (Graphical User Interface)での作業は、デスクトップというフォルダをダブルクリックして、そこを眺めるという流れ

```
コマンド プロンプト
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:¥Users¥kadota>dir
ドライブ C のボリューム ラベルは Windows です
ボリューム シリアル番号は 2083-72DA です

C:¥Users¥kadota のディレクトリ

2016/08/20  15:04  <DIR>          .
2016/08/20  15:04  <DIR>          ..
2016/08/07  08:46  <DIR>          .oracle_jre_u
2016/08/20  11:26  <DIR>          .VirtualBox
2016/06/16  10:18  <DIR>          Contacts
2016/08/18  17:06  <DIR>          Desktop
2016/08/20  10:56  <DIR>          Documents
2016/08/11  15:12  <DIR>          Downloads
2016/07/26  14:54  <DIR>          Favorites
2014/11/14  11:37  <DIR>          Intel
2015/02/24  20:25  <DIR>          Intel.sav
2016/06/16  10:18  <DIR>          Links
2016/06/16  10:18  <DIR>          Music
2016/03/24  11:55  <DIR>          OneDrive
2016/08/19  09:31  <DIR>          Pictures
2015/02/24  20:25  <DIR>          Roaming
2016/06/16  10:18  <DIR>          Saved Games
2016/06/16  10:18  <DIR>          Searches
2015/09/29  16:47  <DIR>          share
```



# GUIとCUI

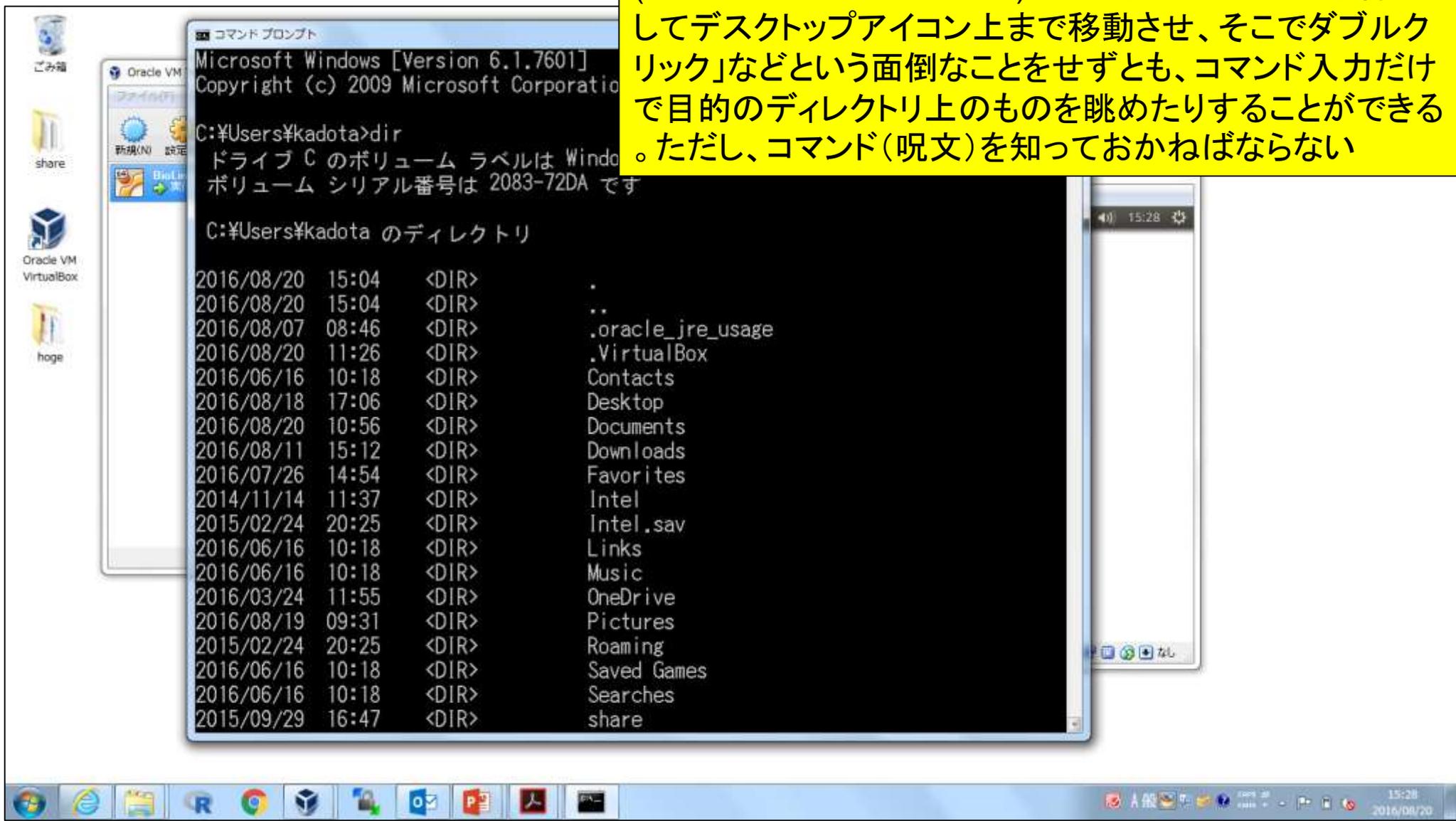
コマンドプロンプト上での作業は、CUI (Character User Interface; Console User Interface)での作業に相当。CLI (Command Line Interface)などともいう。「マウスを操作してデスクトップアイコン上まで移動させ、そこでダブルクリック」などという面倒なことをせずとも、コマンド入力だけで目的のディレクトリ上のものを眺めたりすることができる。ただし、コマンド(呪文)を知っておかねばならない

```
コマンド プロンプト
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation

C:\Users\kadota>dir
ドライブ C のボリューム ラベルは Windows7 です
ボリューム シリアル番号は 2083-72DA です

C:\Users\kadota のディレクトリ

2016/08/20  15:04    <DIR>          .
2016/08/20  15:04    <DIR>          ..
2016/08/07  08:46    <DIR>          .oracle_jre_usage
2016/08/20  11:26    <DIR>          .VirtualBox
2016/06/16  10:18    <DIR>          Contacts
2016/08/18  17:06    <DIR>          Desktop
2016/08/20  10:56    <DIR>          Documents
2016/08/11  15:12    <DIR>          Downloads
2016/07/26  14:54    <DIR>          Favorites
2014/11/14  11:37    <DIR>          Intel
2015/02/24  20:25    <DIR>          Intel.sav
2016/06/16  10:18    <DIR>          Links
2016/06/16  10:18    <DIR>          Music
2016/03/24  11:55    <DIR>          OneDrive
2016/08/19  09:31    <DIR>          Pictures
2015/02/24  20:25    <DIR>          Roaming
2016/06/16  10:18    <DIR>          Saved Games
2016/06/16  10:18    <DIR>          Searches
2015/09/29  16:47    <DIR>          share
```



# dir Desktop

例えば、コマンドプロンプト起動直後の場所(ホームディレクトリという)は、この場合「C:¥Users¥kadota」に相当する。この場所にいながらにして、Desktop上のものを調べることができる。そのやり方の1つは、①「dir Desktop」

```
コマンド プロンプト
2016/06/16 10:18 <DIR> Links
2016/06/16 10:18 <DIR> Music
2016/03/24 11:55 <DIR> OneDrive
2016/08/19 09:31 <DIR> Pictures
2015/02/24 20:25 <DIR> Roaming
2016/06/16 10:18 <DIR> Saved Games
2016/06/16 10:18 <DIR> Searches
2015/09/29 16:47 <DIR> share
2016/06/16 10:18 <DIR> Videos
2016/08/19 17:21 <DIR> VirtualBox VMs
0 個のファイル 0 バイト
21 個のディレクトリ 254,330,654,720 バイトの空き領域

C:¥Users¥kadota>dir Desktop ①
ドライブ C のボリューム ラベルは Windows です
ボリューム シリアル番号は 2083-72DA です

C:¥Users¥kadota¥Desktop のディレクトリ

2016/08/18 17:06 <DIR> .
2016/08/18 17:06 <DIR> ..
2016/08/18 11:36 <DIR> hoge
2015/11/19 17:30 1,101 Oracle VM VirtualBox.lnk
2016/07/11 10:45 <DIR> share

1 個のファイル 1,101 バイト
4 個のディレクトリ 254,324,953,088 バイトの空き領域

C:¥Users¥kadota>
```



もう1つのやり方は、②「cd Desktop」で作業ディレクトリをDesktopに移動してから...

# cd Desktop

```
コマンド プロンプト
2016/03/24 11:55 <DIR> OneDrive
2016/08/19 09:31 <DIR> Pictures
2015/02/24 20:25 <DIR> Roaming
2016/06/16 10:18 <DIR> Saved Games
2016/06/16 10:18 <DIR> Searches
2015/09/29 16:47 <DIR> share
2016/06/16 10:18 <DIR> Videos
2016/08/19 17:21 <DIR> VirtualBox VMs
0 個のファイル 0 バイト
21 個のディレクトリ 254,330,654,720 バイトの空き領域

C:¥Users¥kadota>dir Desktop ①
ドライブ C のボリューム ラベルは Windows です
ボリューム シリアル番号は 2083-72DA です

C:¥Users¥kadota¥Desktop のディレクトリ

2016/08/18 17:06 <DIR> .
2016/08/18 17:06 <DIR> ..
2016/08/18 11:36 <DIR> hoge
2015/11/19 17:30 1,101 Oracle VM VirtualBox.lnk
2016/07/11 10:45 <DIR> share
1 個のファイル 1,101 バイト
4 個のディレクトリ 254,324,953,088 バイトの空き領域

C:¥Users¥kadota>cd Desktop ②
C:¥Users¥kadota¥Desktop>
```

# dir

もう1つのやり方は、②「cd Desktop」で作業ディレクトリをDesktopに移動してから...  
③「dir」。確かに④同じ結果になっている

```
コマンド プロンプト
C:¥Users¥kadota¥Desktop のディレクトリ
2016/08/18 17:06 <DIR> .
2016/08/18 17:06 <DIR> ..
2016/08/18 11:36 <DIR> hoge
2015/11/19 17:30 1,101 Oracle VM VirtualBox.lnk
2016/07/11 10:45 <DIR> share
1 個のファイル 1,101 バイト
4 個のディレクトリ 254,324,953,088 バイトの空き領域

C:¥Users¥kadota>cd Desktop
C:¥Users¥kadota¥Desktop>dir
ドライブ C のボリューム ラベルは Windows です
ボリューム シリアル番号は 2083-72DA です

C:¥Users¥kadota¥Desktop のディレクトリ
2016/08/18 17:06 <DIR> .
2016/08/18 17:06 <DIR> ..
2016/08/18 11:36 <DIR> hoge
2015/11/19 17:30 1,101 Oracle VM VirtualBox.lnk
2016/07/11 10:45 <DIR> share
1 個のファイル 1,101 バイト
4 個のディレクトリ 254,289,620,992 バイトの空き領域

C:¥Users¥kadota¥Desktop>
```

# 作業ディレクトリの把握

②「cd Desktop」実行前後で、赤下線部分が変化していることがわかる。つまり、この部分を眺めることで、今自分がどこで作業をしているかがわかる

```
コマンド プロンプト
C:¥Users¥kadota¥Desktop のディレクトリ
2016/08/18 17:06 <DIR> .
2016/08/18 17:06 <DIR> ..
2016/08/18 11:36 <DIR> hoge
2015/11/19 17:30 1,101 Oracle VM VirtualBox.lnk
2016/07/11 10:45 <DIR> share
1 個のファイル 1,101 バイト
4 個のディレクトリ 254,324,953,088 バイトの空き領域

C:¥Users¥kadota>cd Desktop
C:¥Users¥kadota¥Desktop>dir
ドライブ C のボリューム ラベルは Windows です
ボリューム シリアル番号は 2083-72DA です

C:¥Users¥kadota¥Desktop のディレクトリ
2016/08/18 17:06 <DIR> .
2016/08/18 17:06 <DIR> ..
2016/08/18 11:36 <DIR> hoge
2015/11/19 17:30 1,101 Oracle VM VirtualBox.lnk
2016/07/11 10:45 <DIR> share
1 個のファイル 1,101 バイト
4 個のディレクトリ 254,289,620,992 バイトの空き領域

C:¥Users¥kadota¥Desktop>
```



# Contents

## ■ イン트로ダクション

- 概要、背景(NGS用カリキュラム、講習会)、Linuxスキル習得の意義
- ウェブ情報(日本乳酸菌学会誌のNGS連載やNGS講習会資料)

## ■ 実習環境に慣れる

- 仮想環境での作業に慣れる
- GUIとCUI(マウス操作かコマンド入力操作か)
- ターミナルでの作業
- 共有フォルダの概念を理解

## ■ 練習

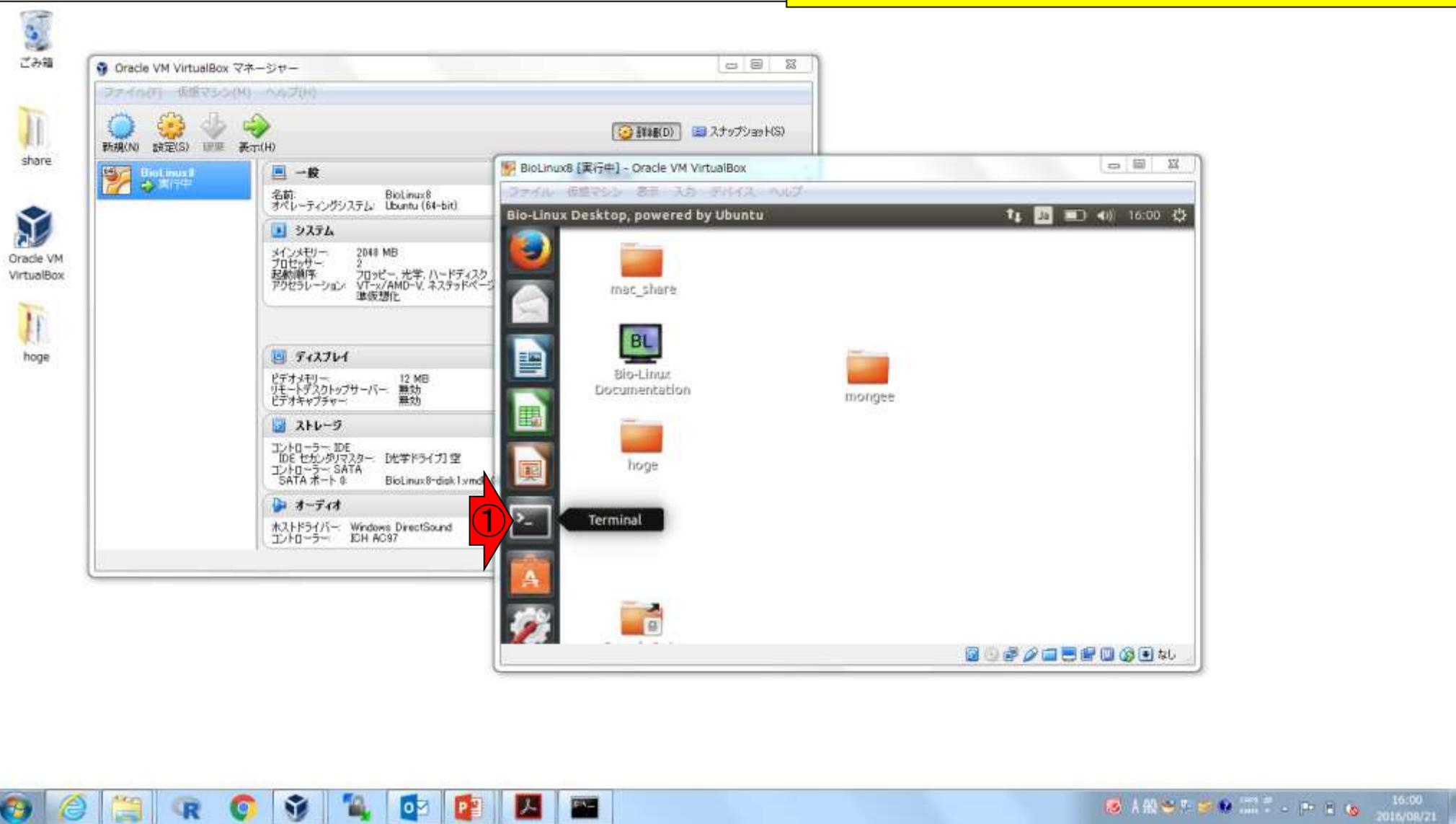
- 作業ディレクトリの変更、練習用NGSデータファイルのダウンロード
- ファイルの確認、de novoゲノムアセンブリ
- BLAST検索

## ■ 課題

- グループごとに異なる課題ファイルを入力として、「ダウンロード、de novoアセンブリ、BLAST検索」を実行し、得られた結果をレポートにまとめて発表せよ
- グループ1はkadai1.fasta、グループ2はkadai2.fasta、etc.

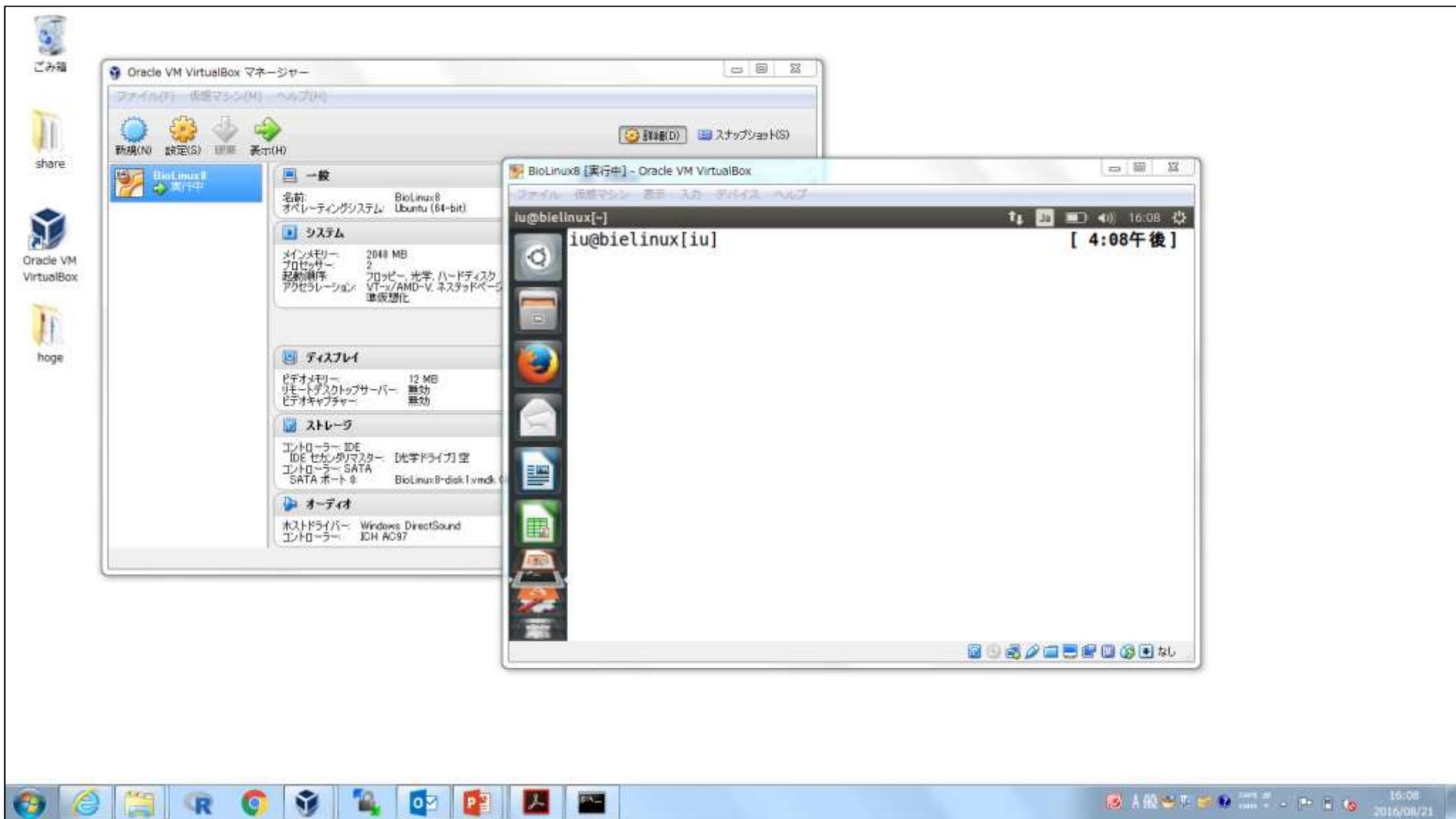
# ターミナル

Windowsのコマンドプロンプトに対応するものは、Linuxでは(Macintosh同様)①ターミナル。第3回ウェブ資料(W8-3;スライド50)あたり



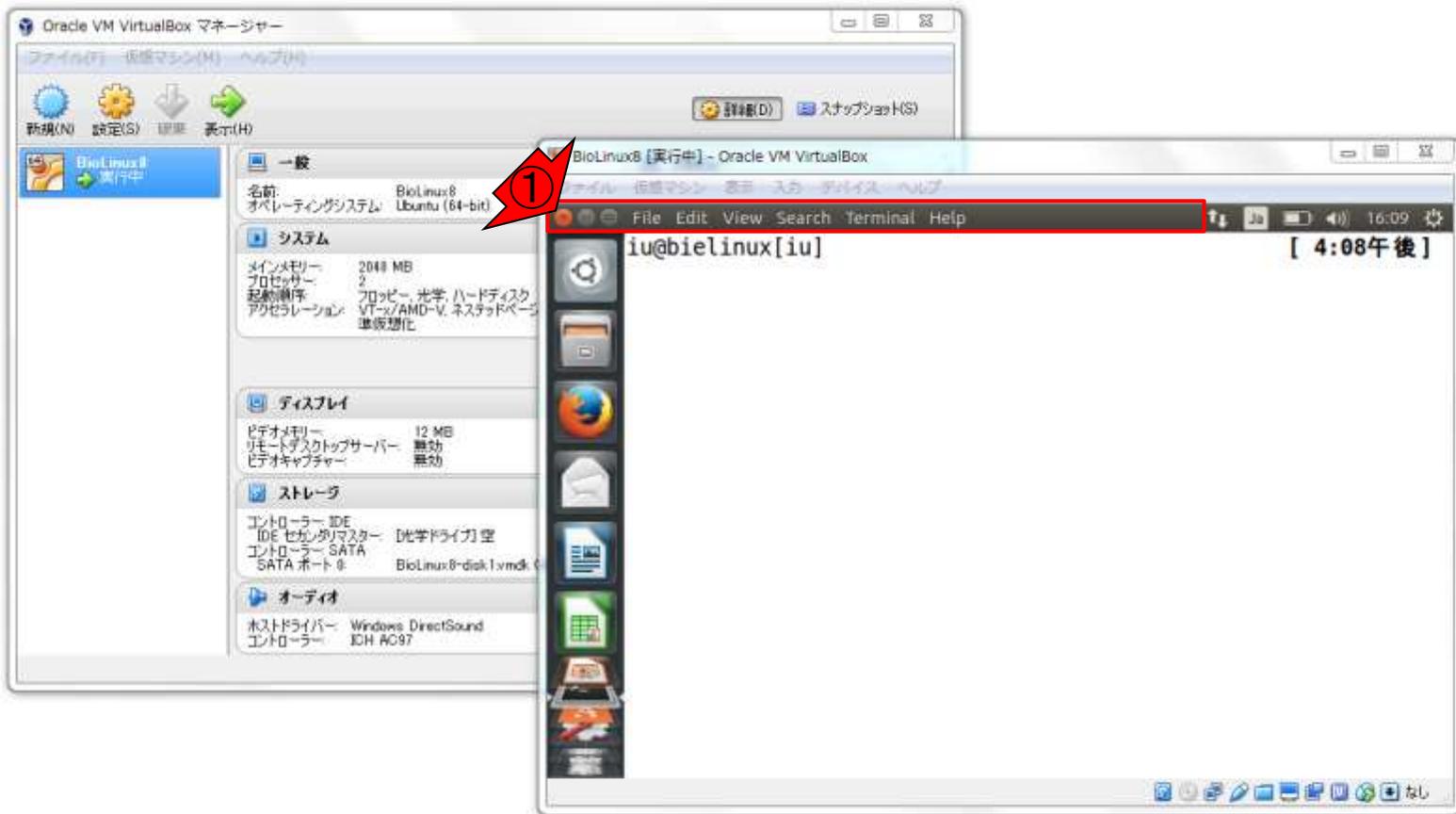
# ターミナル

こんな感じになります。これはターミナルがLinux画面いっぱいになっている状態です



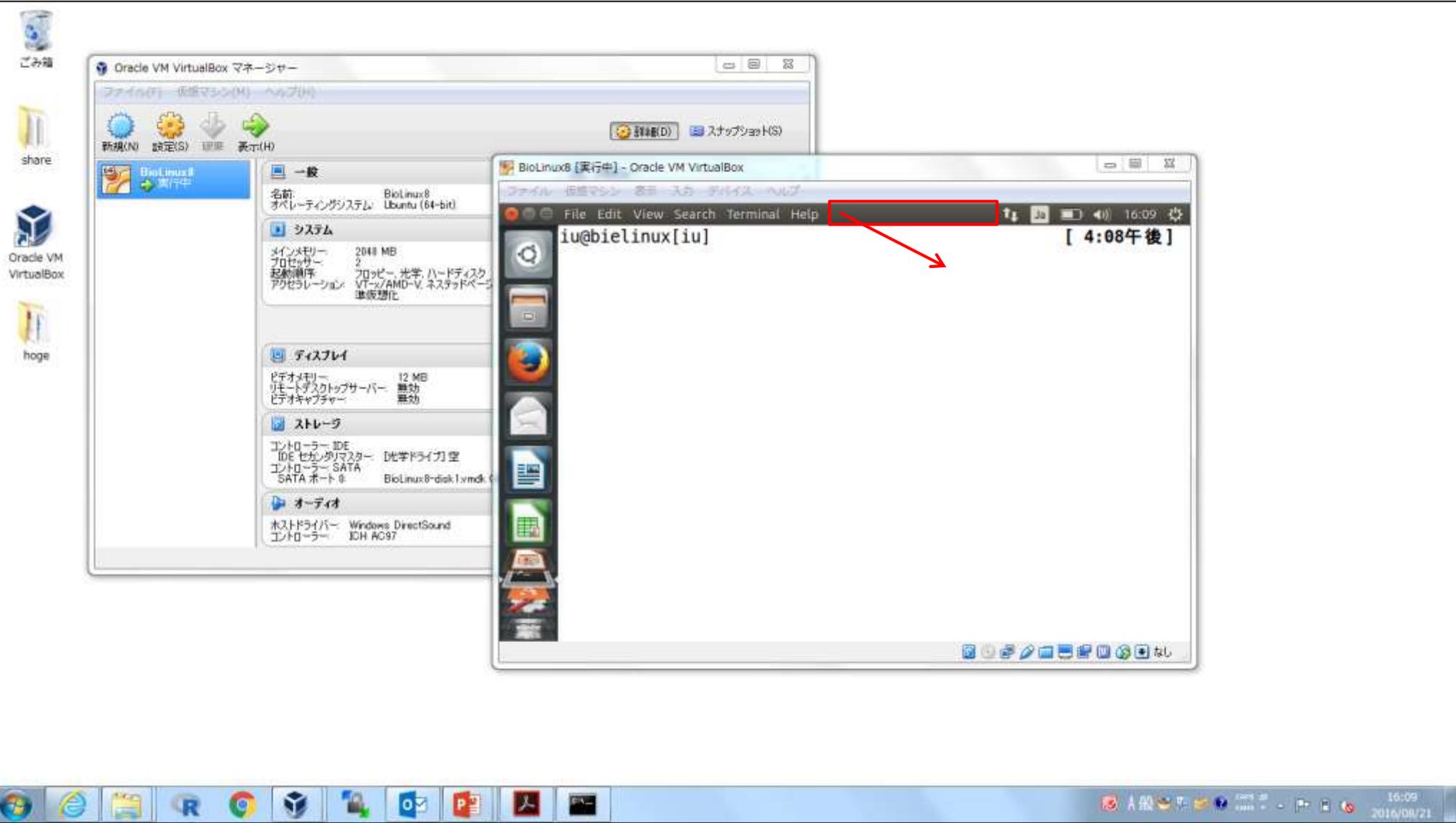
# ターミナル

赤枠あたりにカーソルをもっていくと、メニューバーが見られます。①一番左の×ボタンを押すと、ターミナルを終了できます(が押さない)



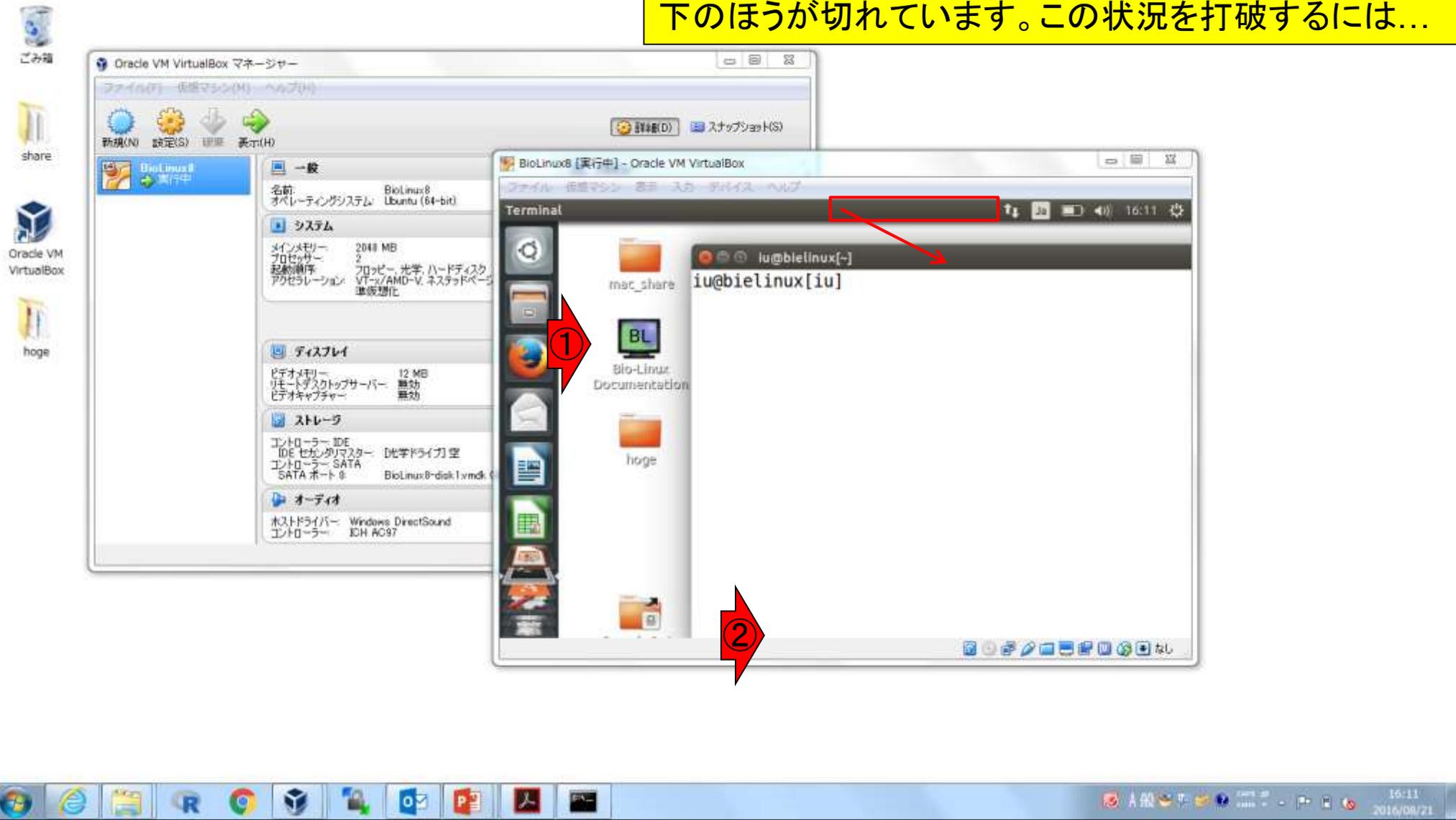
# ターミナル

赤枠あたりで、矢印の始点から終点に向かってドラッグ&ドロップすると...



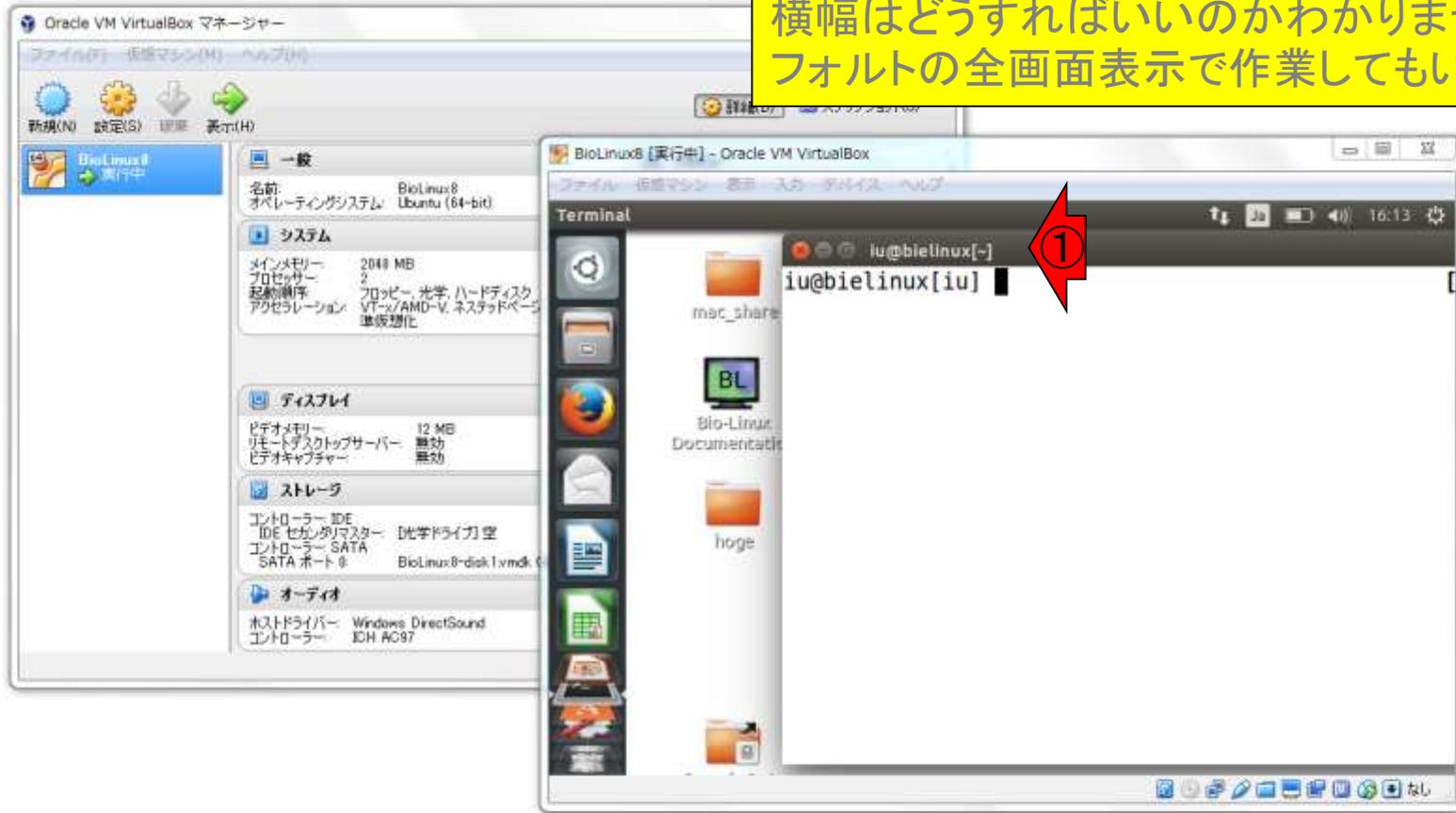
# ターミナル

こんな感じになって、ターミナル画面を移動させることができます。そのおかげでデスクトップ画面上の①アイコンも見えるようになります。しかし、②ターミナル画面の下のほうが切れています。この状況を打破するには...



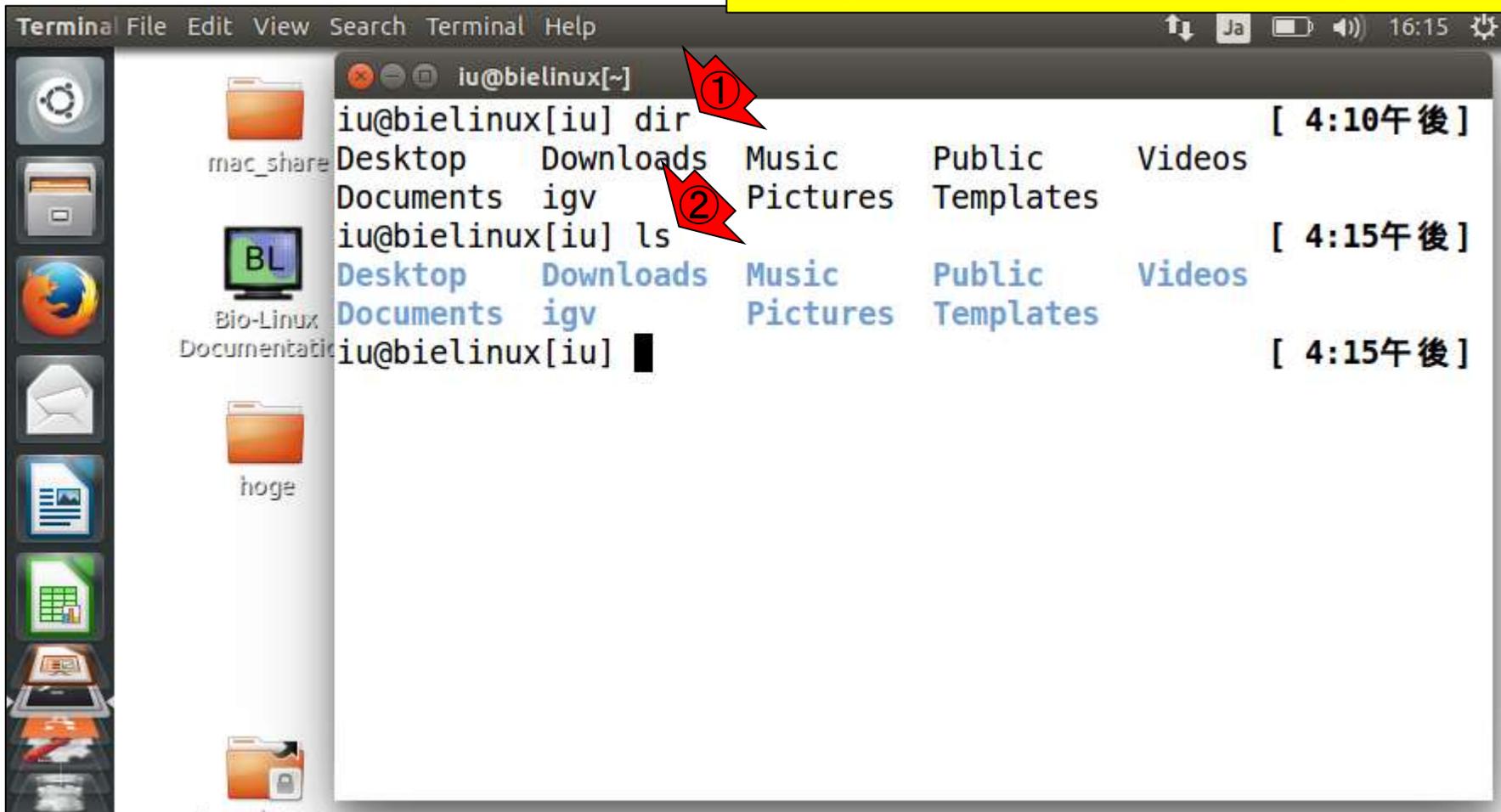
# ターミナル

ターミナル画面の縦幅をLinux画面内に収めるためには、通常はターミナル画面の右下あたりで調整しますが、右下部分が見えていません。①を持って、あちこち動かしていると縦幅をLinux画面内に収めてくれます。横幅はどうすればいいのかわかりませんが、例えばデフォルトの全画面表示で作業してもいいと思います



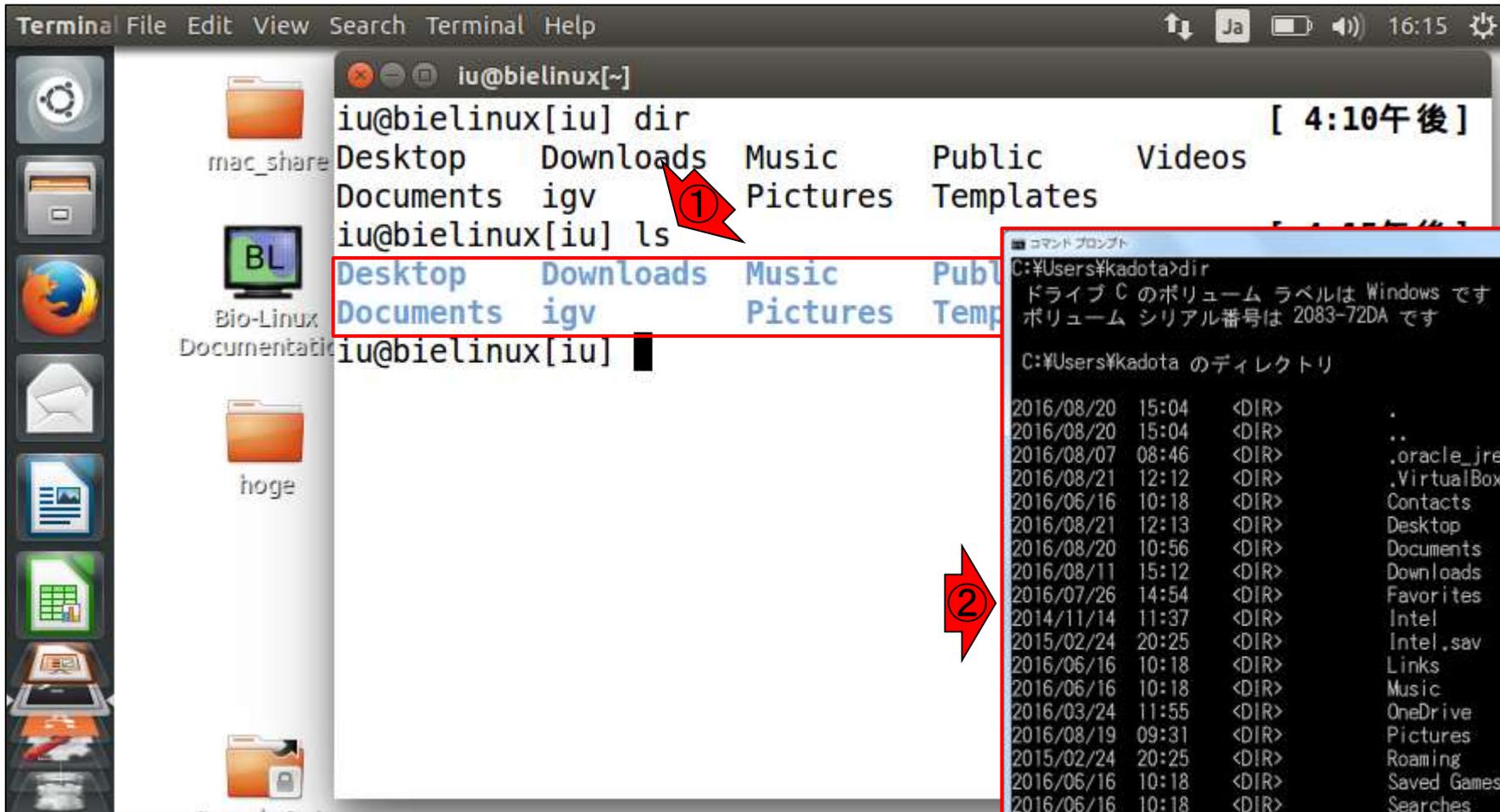
# dirではなくls

作業ディレクトリの中身を確認するのに、Windowsのコマンドプロンプト上では、dirと打ち込みました。Linux環境でも一応①dirで動作しますが、通常は②ls(えるえす)と打つ

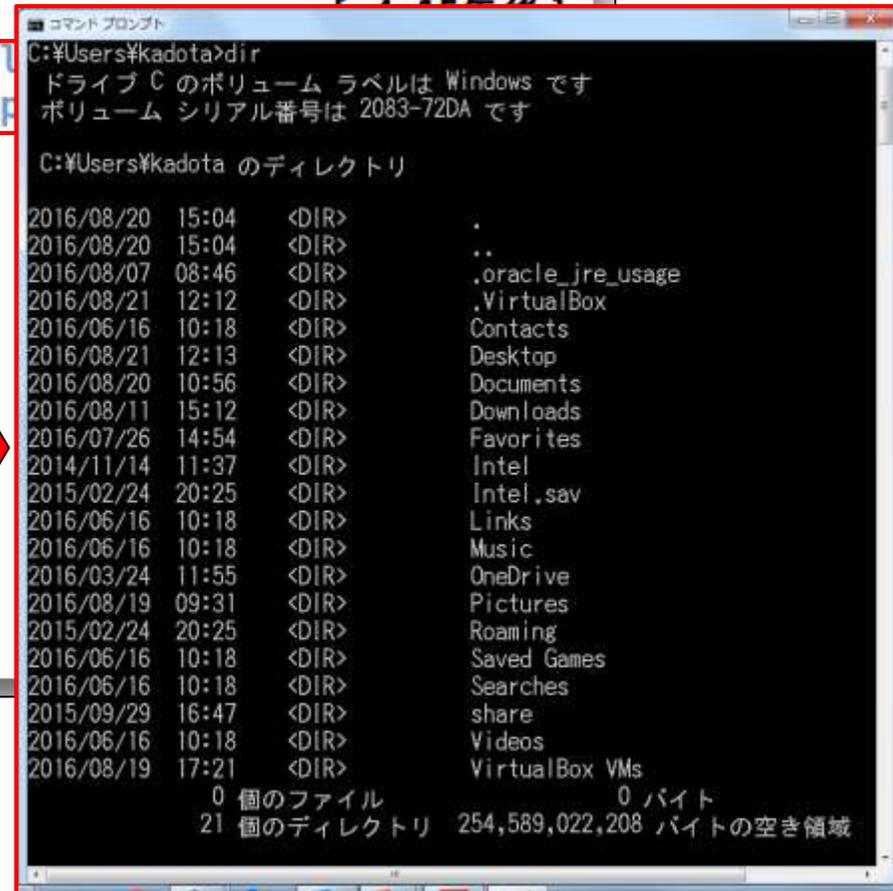


Linux(ホストOS)上での①ls実行結果は、②Windows上でのdir実行結果と似たような感じであることがわかります

# ls実行結果



```
iu@bielinux[iu] dir
Desktop Downloads Music Public Videos
Documents igv Pictures Templates
iu@bielinux[iu] ls
Desktop Downloads Music Pub
Documents igv Pictures Temp
iu@bielinux[iu]
```



```
C:\Users\%kadota>dir
ドライブ C のボリューム ラベルは Windows です
ボリューム シリアル番号は 2083-72DA です

C:\Users\%kadota のディレクトリ

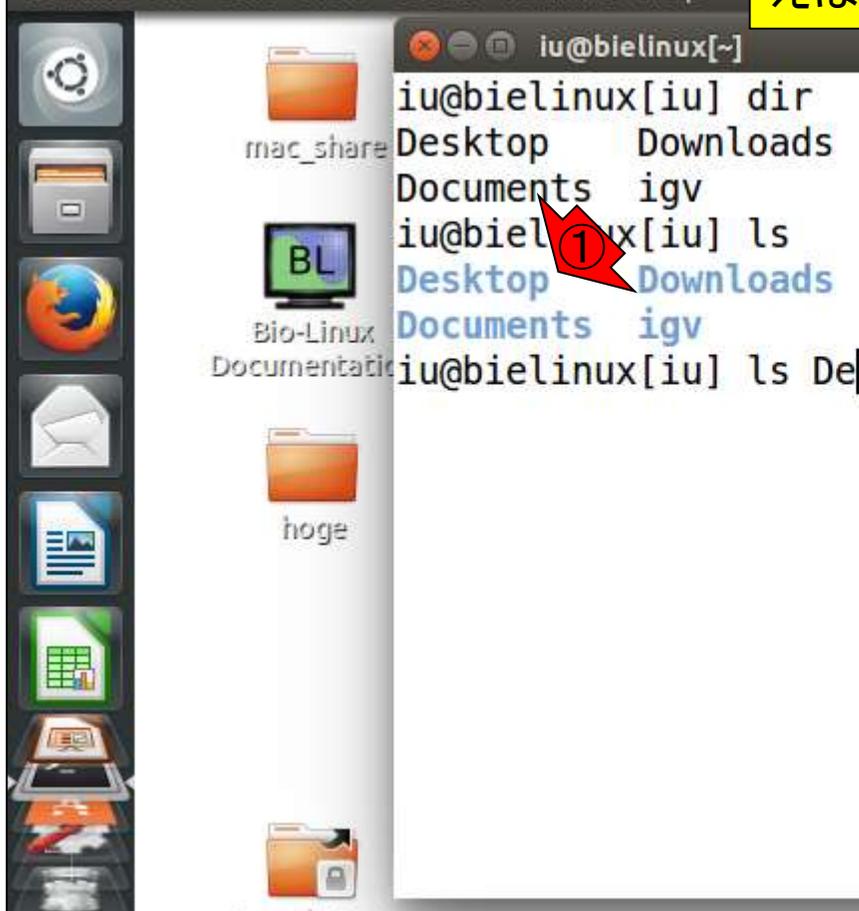
2016/08/20 15:04 <DIR> .
2016/08/20 15:04 <DIR> ..
2016/08/07 08:46 <DIR> .oracle_jre_usage
2016/08/21 12:12 <DIR> .VirtualBox
2016/06/16 10:18 <DIR> Contacts
2016/08/21 12:13 <DIR> Desktop
2016/08/20 10:56 <DIR> Documents
2016/08/11 15:12 <DIR> Downloads
2016/07/26 14:54 <DIR> Favorites
2014/11/14 11:37 <DIR> Intel
2015/02/24 20:25 <DIR> Intel.sav
2016/06/16 10:18 <DIR> Links
2016/06/16 10:18 <DIR> Music
2016/03/24 11:55 <DIR> OneDrive
2016/08/19 09:31 <DIR> Pictures
2015/02/24 20:25 <DIR> Roaming
2016/06/16 10:18 <DIR> Saved Games
2016/06/16 10:18 <DIR> Searches
2015/09/29 16:47 <DIR> share
2016/06/16 10:18 <DIR> Videos
2016/08/19 17:21 <DIR> VirtualBox VMs

0 個のファイル 0 バイト
21 個のディレクトリ 254,589,022,208 バイトの空き領域
```

# ls Desktop

①Desktopというディレクトリが見えているので、その中身を表示させます。「ls Desktop」と打てばいいですが、Linuxの世界では、必要最小限の労力でコマンドを入力するのが基本です。例えば②「ls De」まで打ってから、③Tabキーを押してみましょう

Terminal File Edit View Search Terminal Help

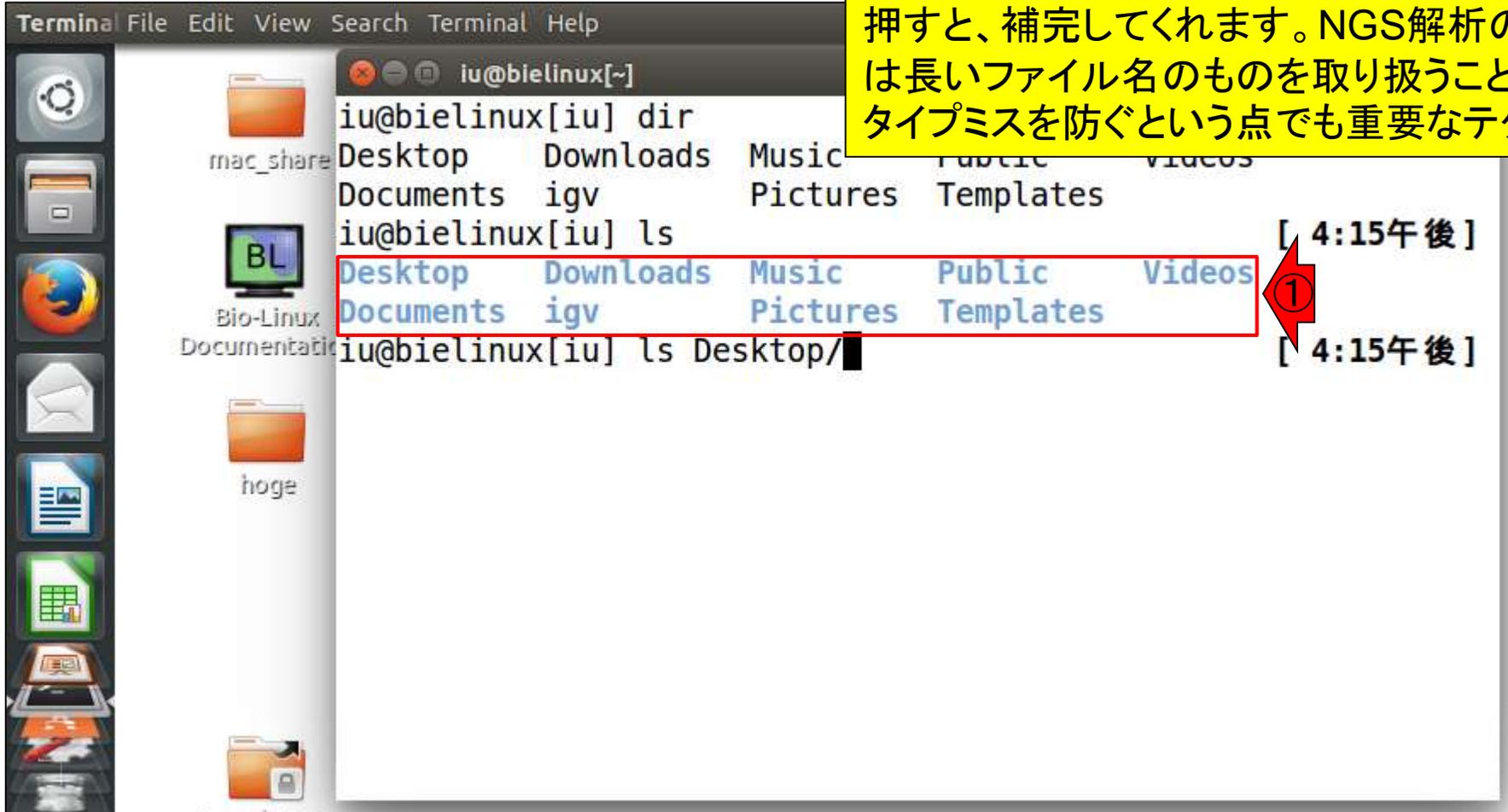


```
iu@bielinux[~]
iu@bielinux[iu] dir [ 4:10午後]
Desktop Downloads Music Public Videos
Documents igv Pictures Templates
iu@bielinux[iu] ls [ 4:15午後]
Desktop Downloads Music Public Videos
Documents igv Pictures Templates
iu@bielinux[iu] ls De [ 4:15午後]
```



# タブ補完

「ls Desktop/」となります。このテクニックを「タブ補完」などと呼ぶ。①赤枠を眺めると、Deから始まるものはDesktopしかない。このような状況でTabキーを押すと、補完してくれます。NGS解析の実務局面では長いファイル名のものを取り扱うこともあるので、タイプミスを防ぐという点でも重要なテクニックです



# ls Desktop

「ls Desktop」実行結果。確かに赤枠で示すように、Linuxのデスクトップ画面に見えているものと同じものが見えている。①mongeeはヒトそれぞれ。ここまでの作業はターミナル起動直後の「ホームディレクトリ」上で行いました

Terminal File Edit View Search Terminal Help

```
iu@bielinux[~]
iu@bielinux[iu] dir [12:42午後]
Desktop Downloads Music Public Videos
Documents igv Pictures Templates
iu@bielinux[iu] ls [12:42午後]
Desktop Downloads Music Public Videos
Documents igv Pictures Templates
iu@bielinux[iu] ls Desktop [12:42午後]
Bio-Linux Documentation hoge mac_share mongee Sample Data [12:42午後]
iu@bielinux[iu] █
```

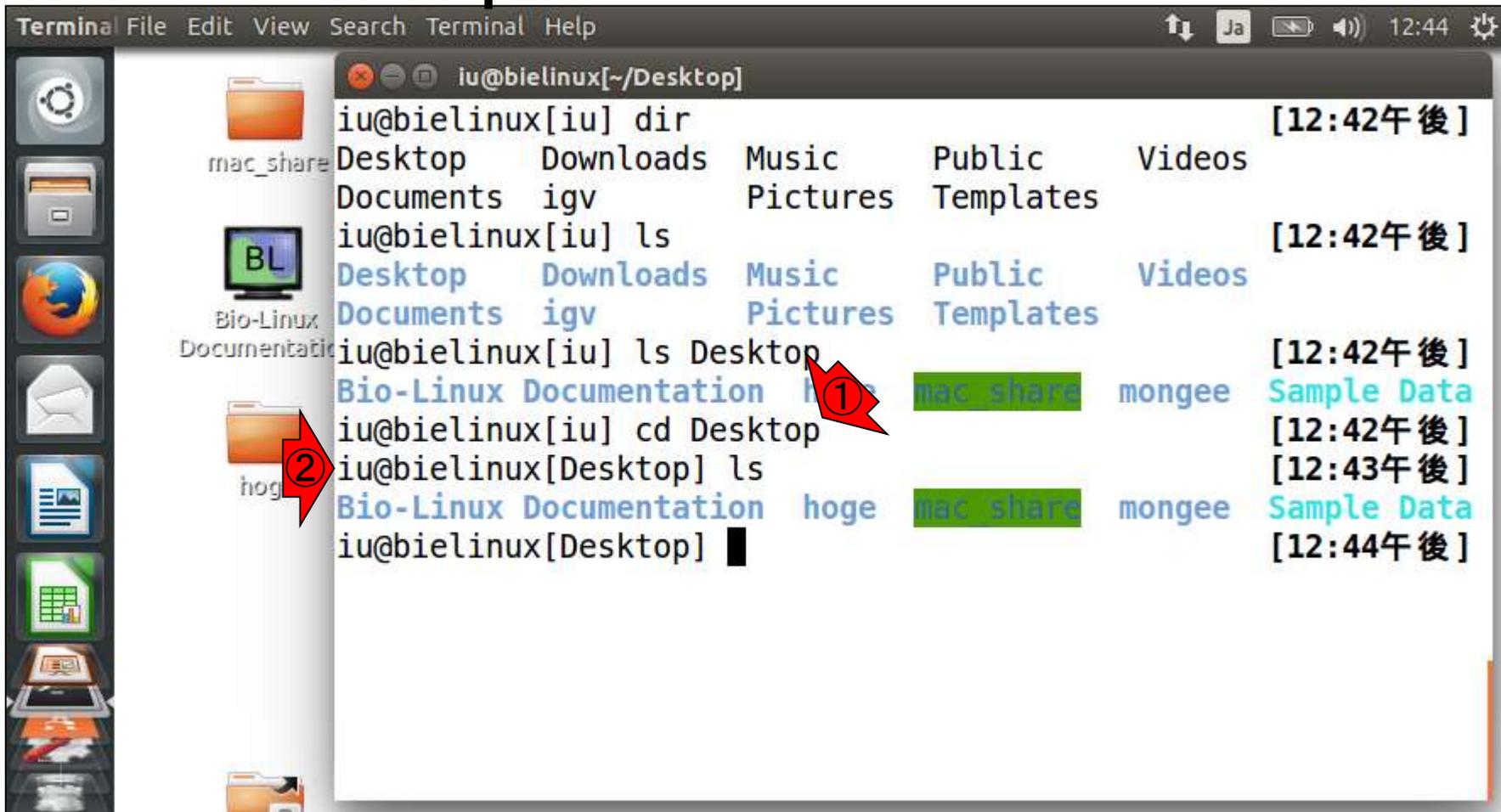
mac\_share

Bio-Linux Documentation

hoge

# cd Desktop

もちろん①cd Desktopとして、Desktopディレクトリに移動してから、②lsするのも構いません



The image shows a terminal window on a Linux desktop environment. The terminal title is 'iu@bielinux[~/Desktop]'. The user performs the following commands and receives the following output:

```
iu@bielinux[iu] dir [12:42午後]
Desktop Downloads Music Public Videos
Documents igv Pictures Templates

iu@bielinux[iu] ls [12:42午後]
Desktop Downloads Music Public Videos
Documents igv Pictures Templates

iu@bielinux[iu] ls Desktop [12:42午後]
Bio-Linux Documentation hoge mac share mongee Sample Data

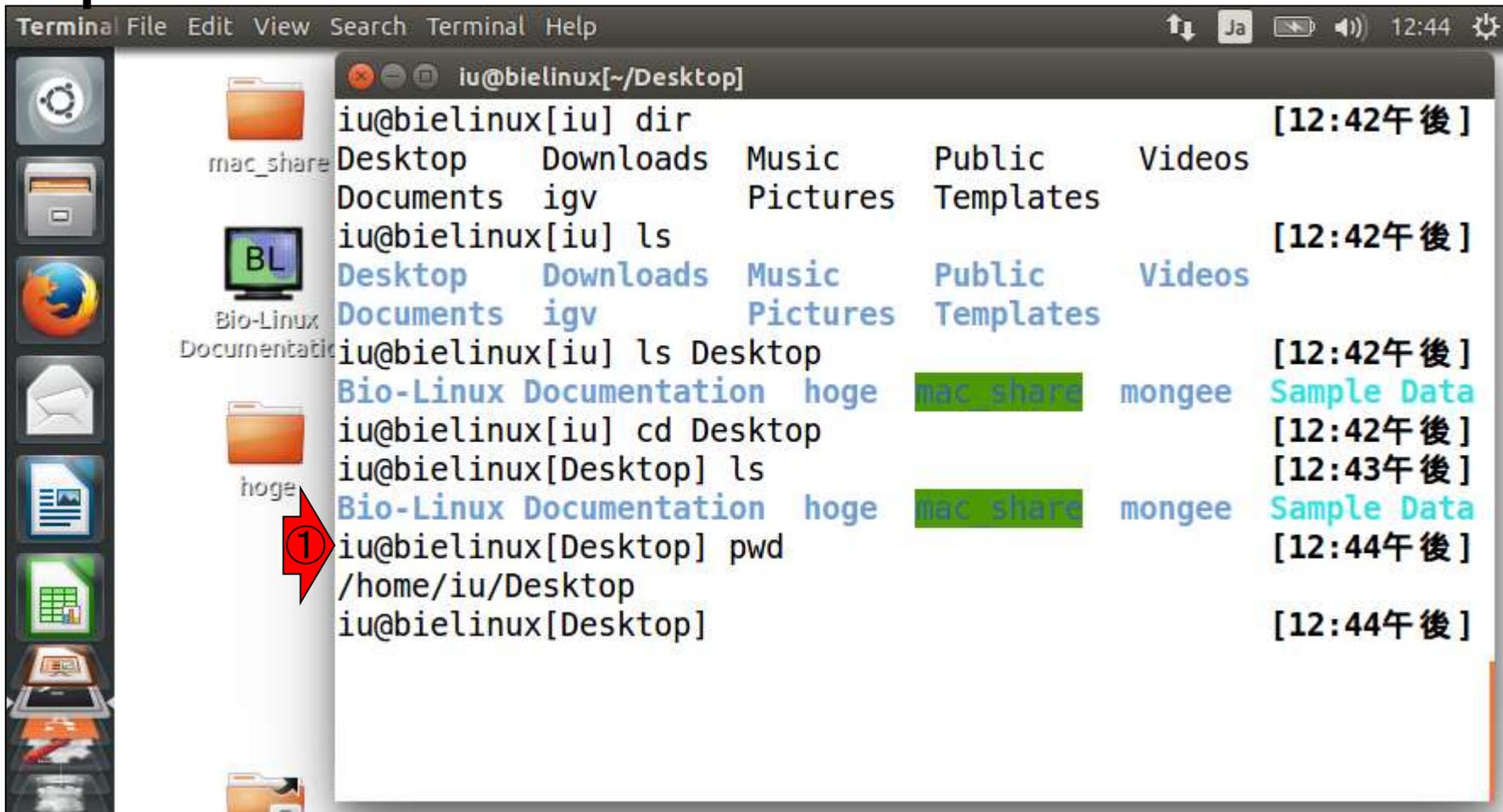
iu@bielinux[iu] cd Desktop [12:42午後]
iu@bielinux[Desktop] ls [12:43午後]
Bio-Linux Documentation hoge mac share mongee Sample Data

iu@bielinux[Desktop] █ [12:44午後]
```

Red arrows point to the 'cd Desktop' command (labeled ①) and the 'ls Desktop' command (labeled ②). The terminal output shows that the 'ls Desktop' command lists the contents of the Desktop directory, including 'mac share' and 'mongee', which are highlighted in green in the original image.

# pwd

①pwdで現在の作業ディレクトリを表示させています(print working directory)



The image shows a terminal window titled "iu@bielinux[~/Desktop]". The terminal output is as follows:

```
iu@bielinux[iu] dir [12:42午後]
Desktop Downloads Music Public Videos
Documents igv Pictures Templates

iu@bielinux[iu] ls [12:42午後]
Desktop Downloads Music Public Videos
Documents igv Pictures Templates

iu@bielinux[iu] ls Desktop [12:42午後]
Bio-Linux Documentation hoge mac share mongee Sample Data

iu@bielinux[iu] cd Desktop [12:42午後]
iu@bielinux[Desktop] ls [12:43午後]
Bio-Linux Documentation hoge mac share mongee Sample Data

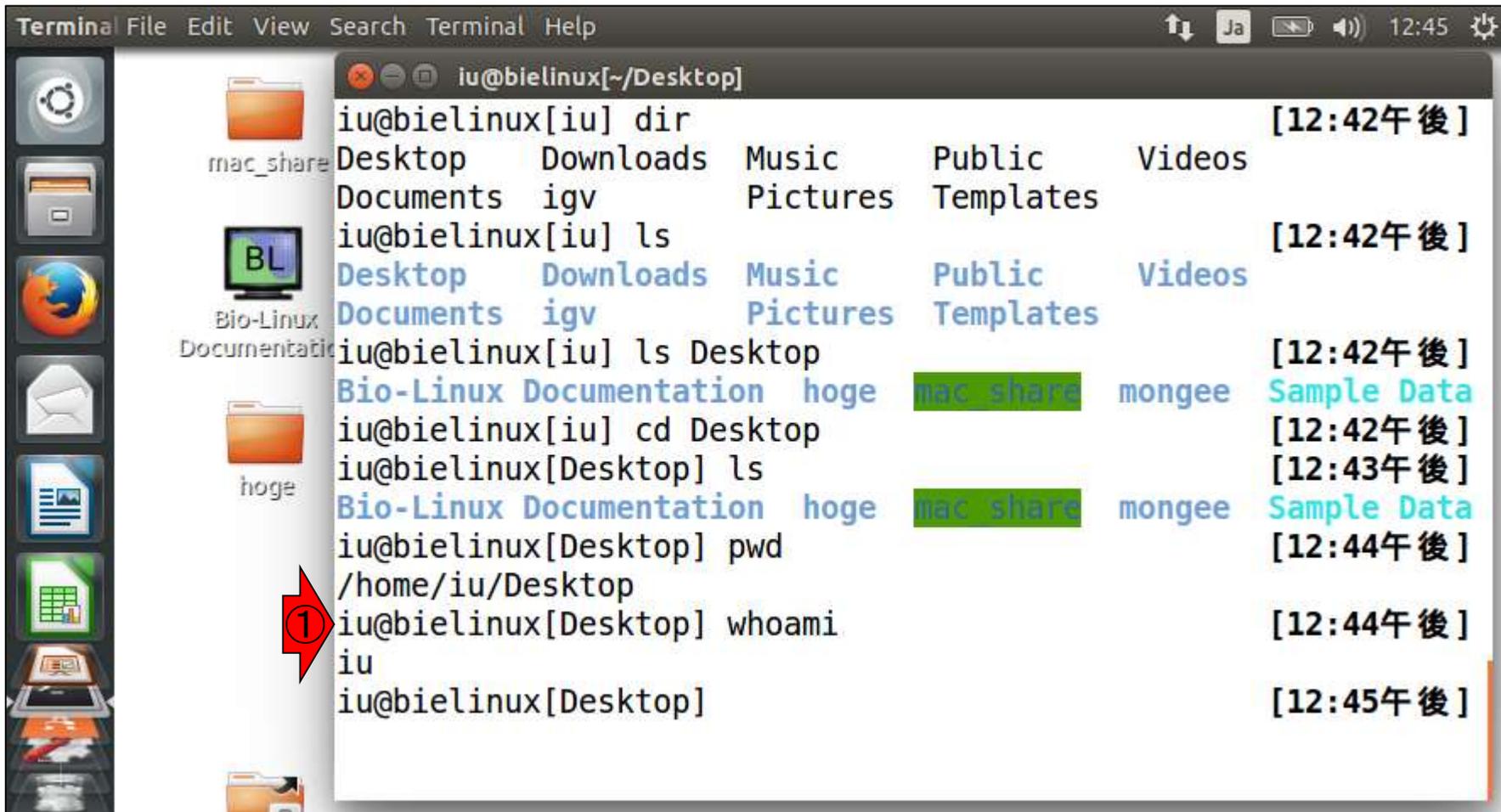
iu@bielinux[Desktop] pwd [12:44午後]
/home/iu/Desktop

iu@bielinux[Desktop] [12:44午後]
```

A red arrow with the number "1" points to the `pwd` command and its output, `/home/iu/Desktop`.

①whoamiでユーザ名(iu)を調べることができます

# whoami



```
Terminal File Edit View Search Terminal Help
iu@bielinux[~/Desktop]
iu@bielinux[iu] dir [12:42午後]
Desktop Downloads Music Public Videos
Documents igv Pictures Templates
iu@bielinux[iu] ls [12:42午後]
Desktop Downloads Music Public Videos
Documents igv Pictures Templates
iu@bielinux[iu] ls Desktop [12:42午後]
Bio-Linux Documentation hoge mac share mongee Sample Data
iu@bielinux[iu] cd Desktop [12:42午後]
iu@bielinux[Desktop] ls [12:43午後]
Bio-Linux Documentation hoge mac share mongee Sample Data
iu@bielinux[Desktop] pwd [12:44午後]
/home/iu/Desktop
① iu@bielinux[Desktop] whoami [12:44午後]
iu
iu@bielinux[Desktop] [12:45午後]
```

# mac\_share

①貸与PCは、mac\_shareというディレクトリが反転されていると思います。macというキーワードから、Macintoshを連想するヒトがいるかもしれませんが、ただの文字列であり無関係です

Terminal File Edit View Search Terminal Help

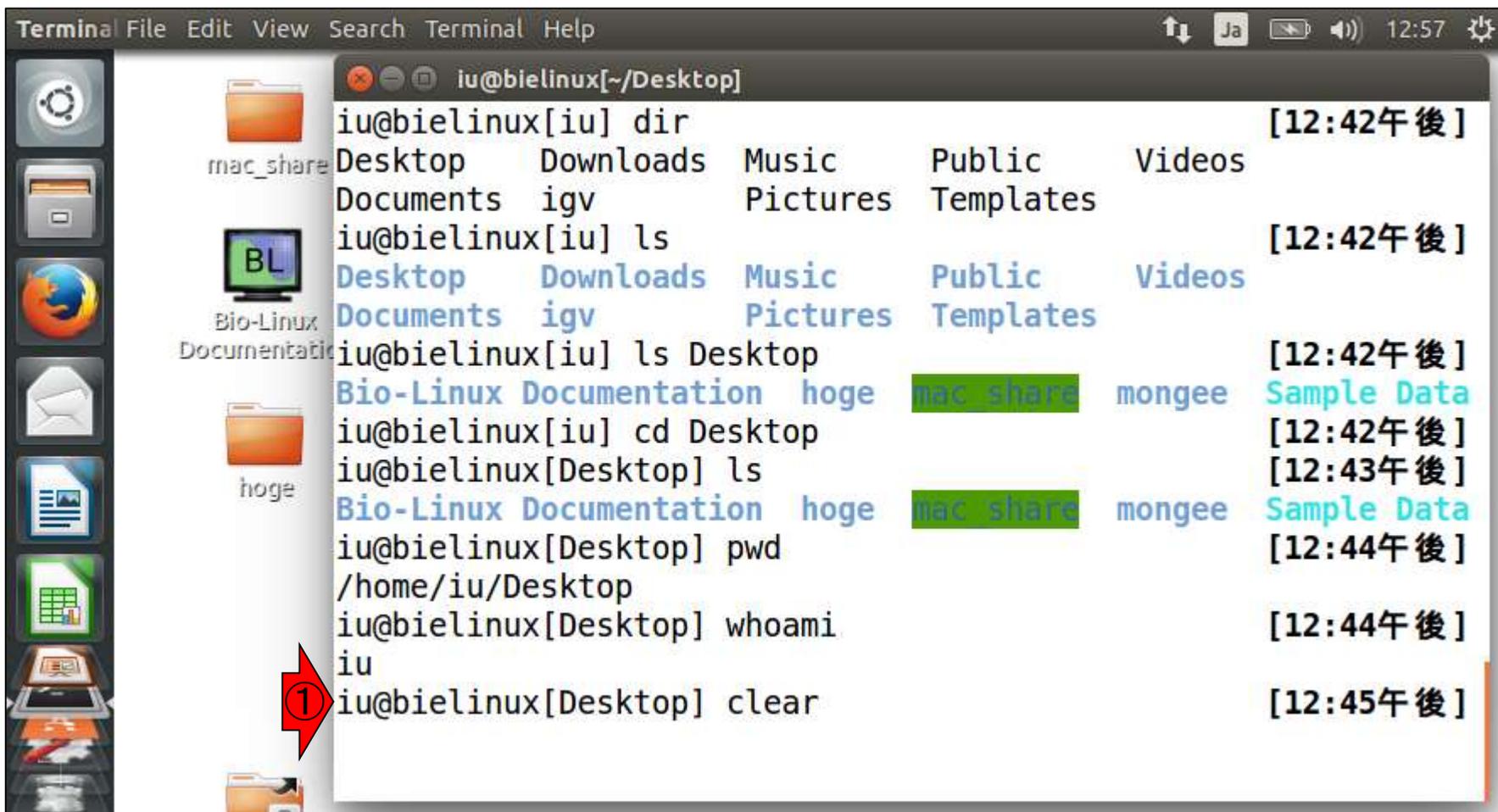
iu@bielinux[~/Desktop]

```
iu@bielinux[iu] dir [12:42午後]
Desktop Downloads Music Public Videos
Documents igv Pictures Templates
iu@bielinux[iu] ls [12:42午後]
Desktop Downloads Music Public Videos
Documents igv Pictures Templates
iu@bielinux[iu] ls Desktop [12:42午後]
Bio-Linux Documentation hoge mac_share mongee Sample Data
iu@bielinux[iu] cd Desktop [12:42午後]
iu@bielinux[Desktop] ls [12:43午後]
Bio-Linux Documentation hoge mac_share mongee Sample Data
iu@bielinux[Desktop] pwd [12:44午後]
/home/iu/Desktop
iu@bielinux[Desktop] whoami [12:44午後]
iu
iu@bielinux[Desktop] [12:45午後]
```

The image shows a Linux desktop environment with a terminal window open. The terminal displays a series of commands and their outputs. A red arrow with the number '1' points to the 'mac\_share' directory in the file manager on the left and to the 'mac\_share' entry in the terminal output. The terminal output shows that the 'mac\_share' directory is located in the Desktop directory and contains files named 'mongee' and 'Sample Data'. The terminal also shows the user's current directory is '/home/iu/Desktop' and the user is 'iu'.

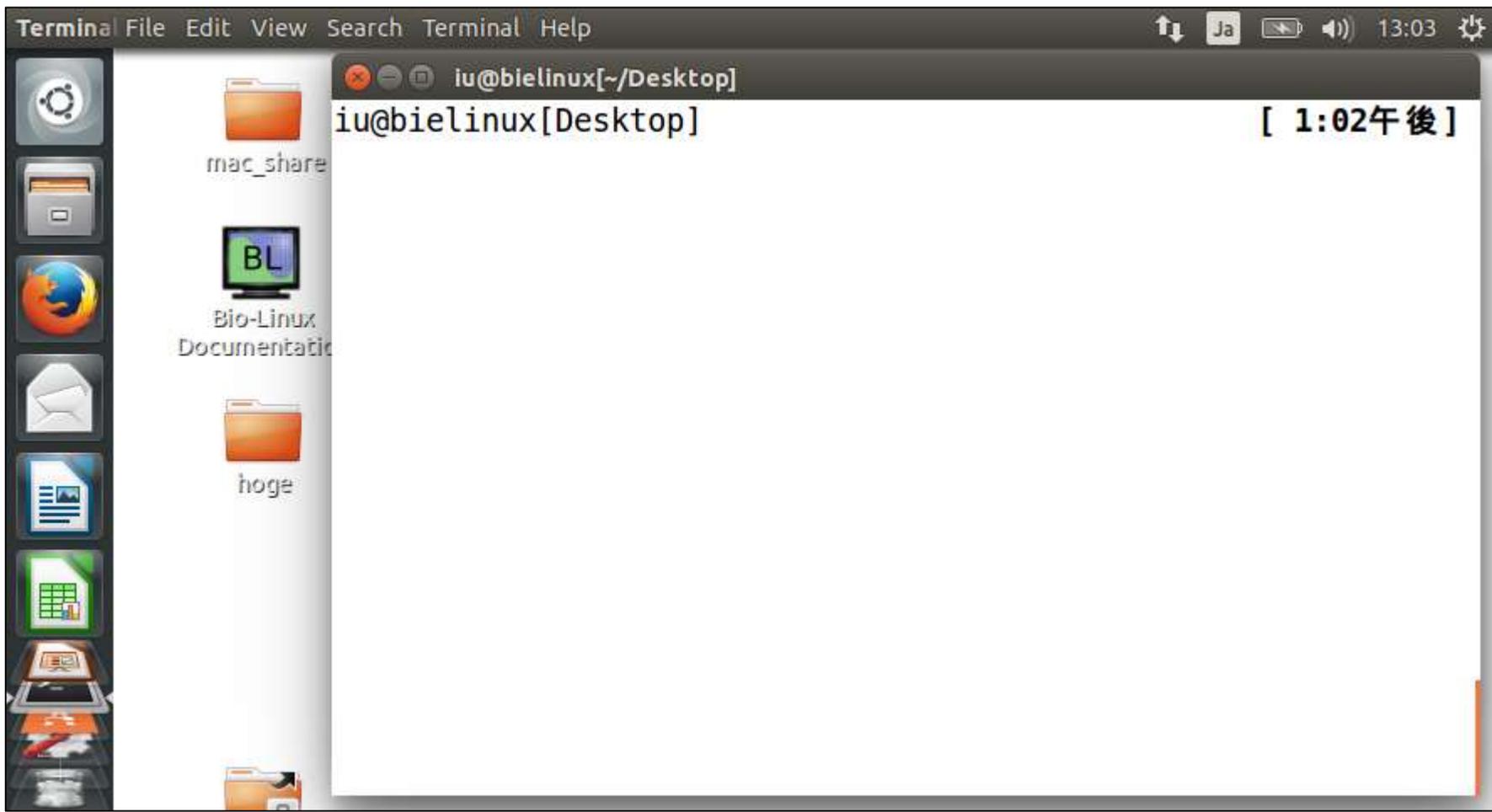
# clear

①clearと打つことで、ターミナル画面をリフレッシュすることができます



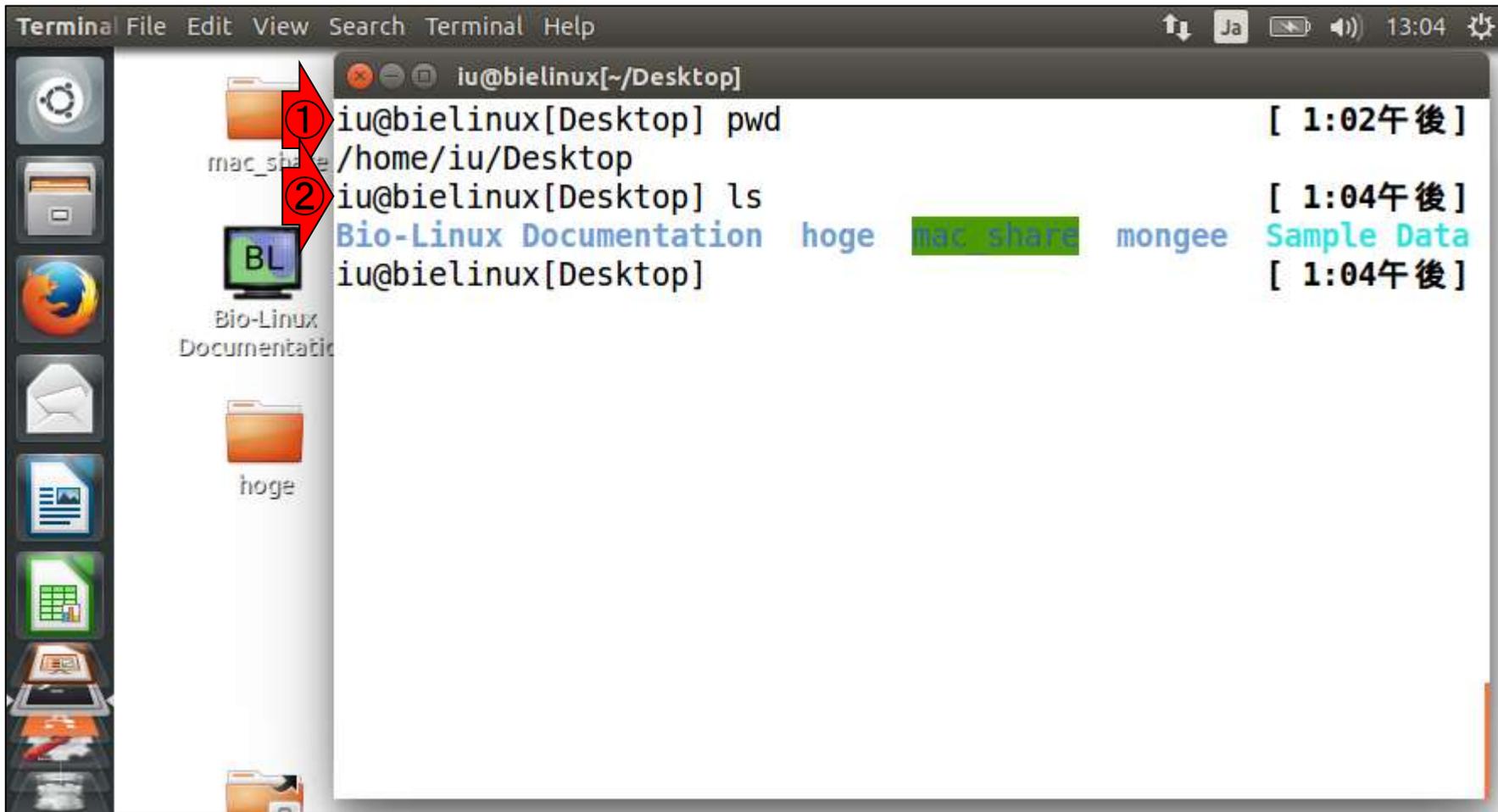
```
Terminal File Edit View Search Terminal Help
iu@bielinux[~/Desktop]
iu@bielinux[iu] dir [12:42午後]
Desktop Downloads Music Public Videos
Documents igv Pictures Templates
iu@bielinux[iu] ls [12:42午後]
Desktop Downloads Music Public Videos
Documents igv Pictures Templates
iu@bielinux[iu] ls Desktop [12:42午後]
Bio-Linux Documentation hoge mac share mongee Sample Data
iu@bielinux[iu] cd Desktop [12:42午後]
iu@bielinux[Desktop] ls [12:43午後]
Bio-Linux Documentation hoge mac share mongee Sample Data
iu@bielinux[Desktop] pwd [12:44午後]
/home/iu/Desktop
iu@bielinux[Desktop] whoami [12:44午後]
iu
iu@bielinux[Desktop] clear [12:45午後]
```

# clear



# clear

①pwd、②ls。作業ディレクトリはclear実行前と同じです



The image shows a Linux desktop environment with a terminal window open. The terminal window title is "iu@bielinux[~/Desktop]". The terminal output is as follows:

```
iu@bielinux[Desktop] pwd [ 1:02午後 ]  
/home/iu/Desktop  
iu@bielinux[Desktop] ls [ 1:04午後 ]  
Bio-Linux Documentation hoge mac share mongee Sample Data  
iu@bielinux[Desktop] [ 1:04午後 ]
```

Red arrows with numbers 1 and 2 point to the first and second lines of the terminal output, respectively. The "mac share" directory name in the ls output is highlighted in green.

# mac\_share

- ① mac\_shareディレクトリに移動して、(pwdで確認し)
- ② ls。このディレクトリ内には何もありません。
- ③ mac\_shareフォルダをダブルクリックして開くと...

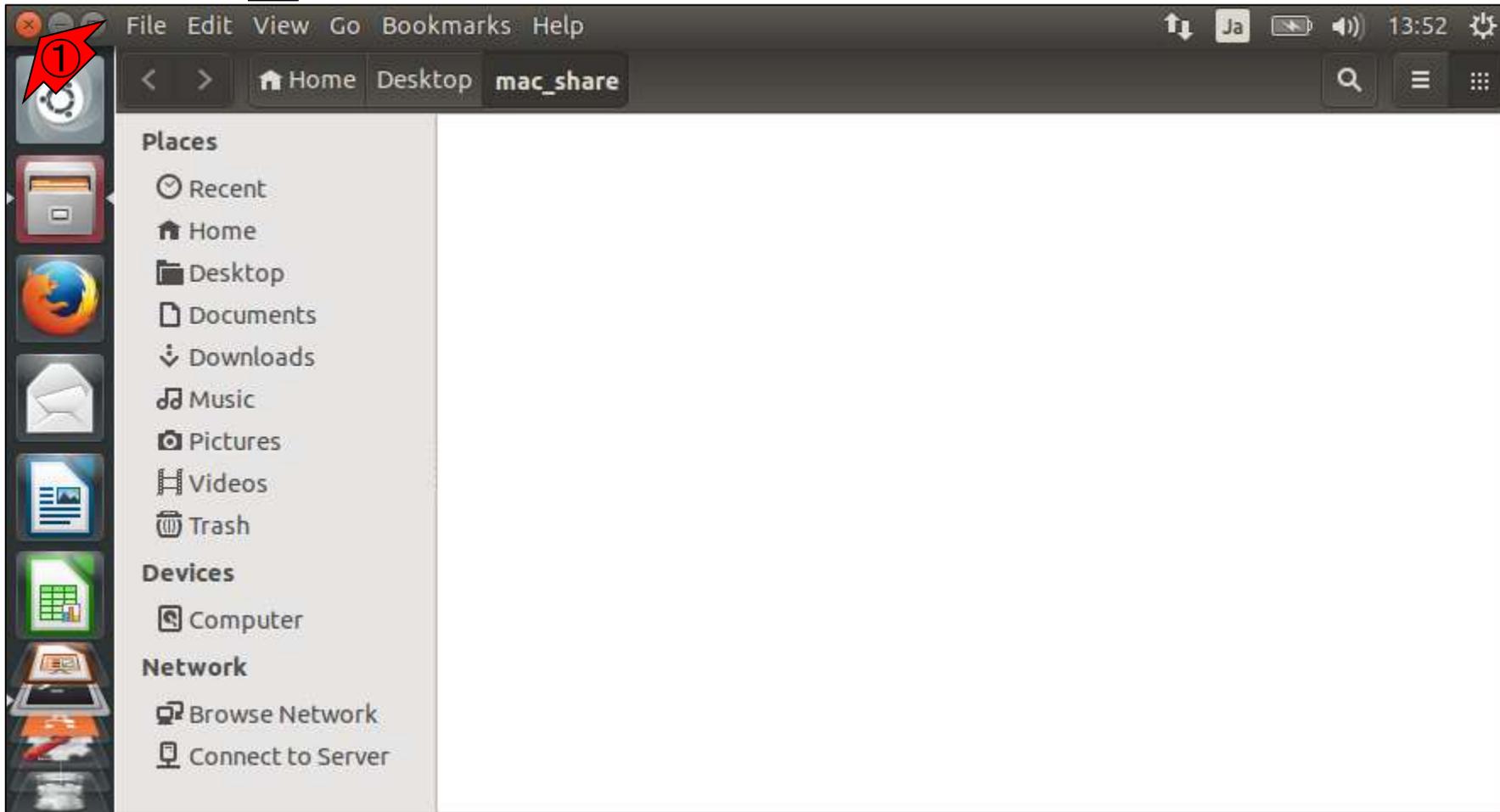
The screenshot shows a Linux desktop environment with a terminal window and a file manager. The terminal window is titled "iu@bielinux[~/Desktop/mac\_share]" and displays the following commands and output:

```
iu@bielinux[Desktop] pwd [ 1:02午後 ]
/home/iu/Desktop
iu@bielinux[Desktop] ls [ 1:04午後 ]
Bio-Linux Documentation hoge mac_share mongee Sample Data
iu@bielinux[Desktop] cd mac_share [ 1:04午後 ]
iu@bielinux[mac_share] pwd [ 1:07午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls [ 1:07午後 ]
iu@bielinux[mac_share] [ 1:08午後 ]
```

The file manager shows the "mac\_share" folder selected, with a red arrow and the number "3" pointing to it. The "Bio-Linux Documentation" folder is also visible, with a red arrow and the number "1" pointing to it. The "hoge" folder is also visible, with a red arrow and the number "2" pointing to it.

# mac\_share

こんな感じになって、このフォルダ中には何も  
ないことがわかります。①×で終了しておく



# Contents

## ■ イン트로ダクション

- 概要、背景(NGS用カリキュラム、講習会)、Linuxスキル習得の意義
- ウェブ情報(日本乳酸菌学会誌のNGS連載やNGS講習会資料)

## ■ 実習環境に慣れる

- 仮想環境での作業に慣れる
- GUIとCUI(マウス操作かコマンド入力操作か)
- ターミナルでの作業
- 共有フォルダの概念を理解

## ■ 練習

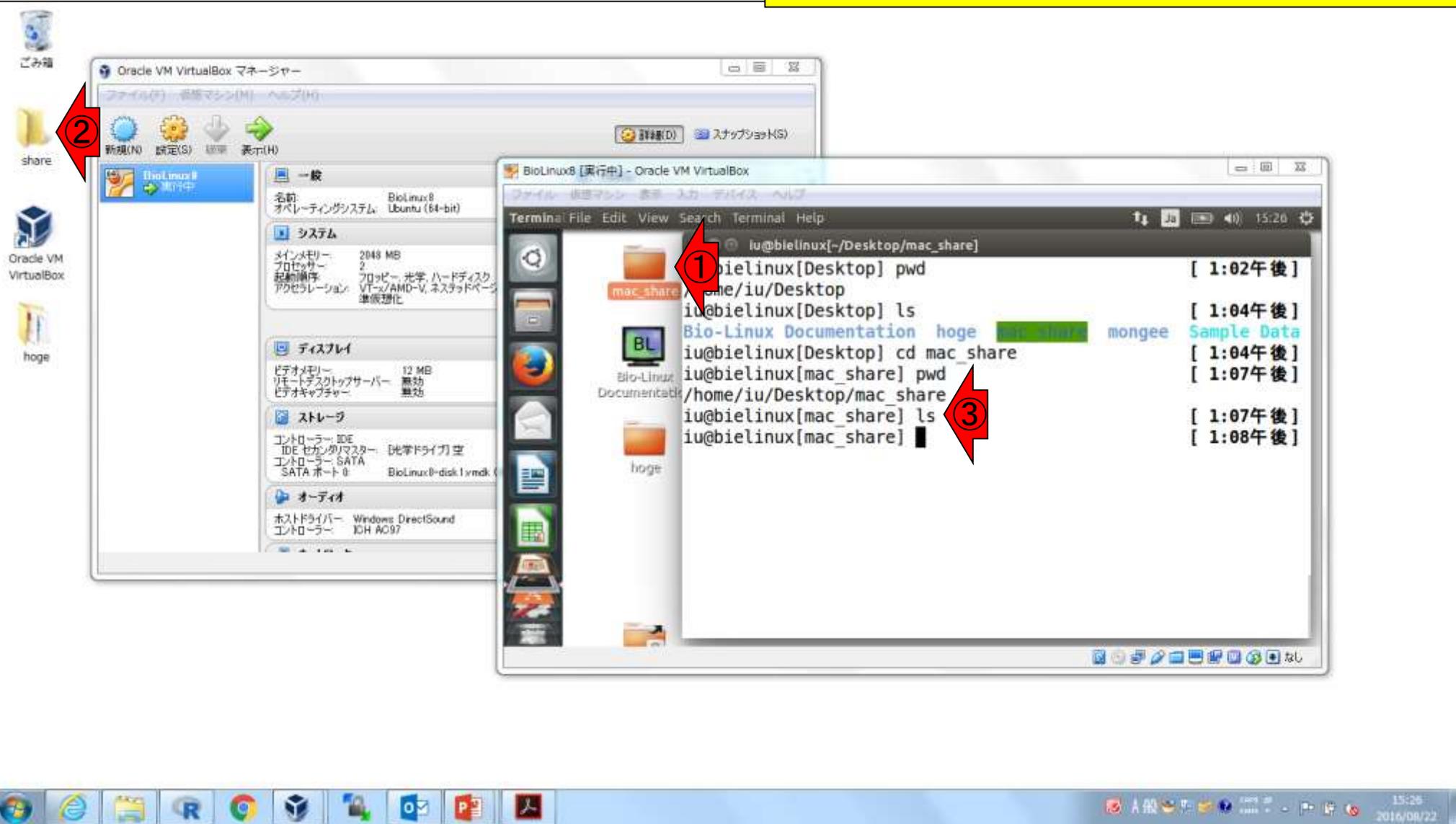
- 作業ディレクトリの変更、練習用NGSデータファイルのダウンロード
- ファイルの確認、de novoゲノムアセンブリ
- BLAST検索

## ■ 課題

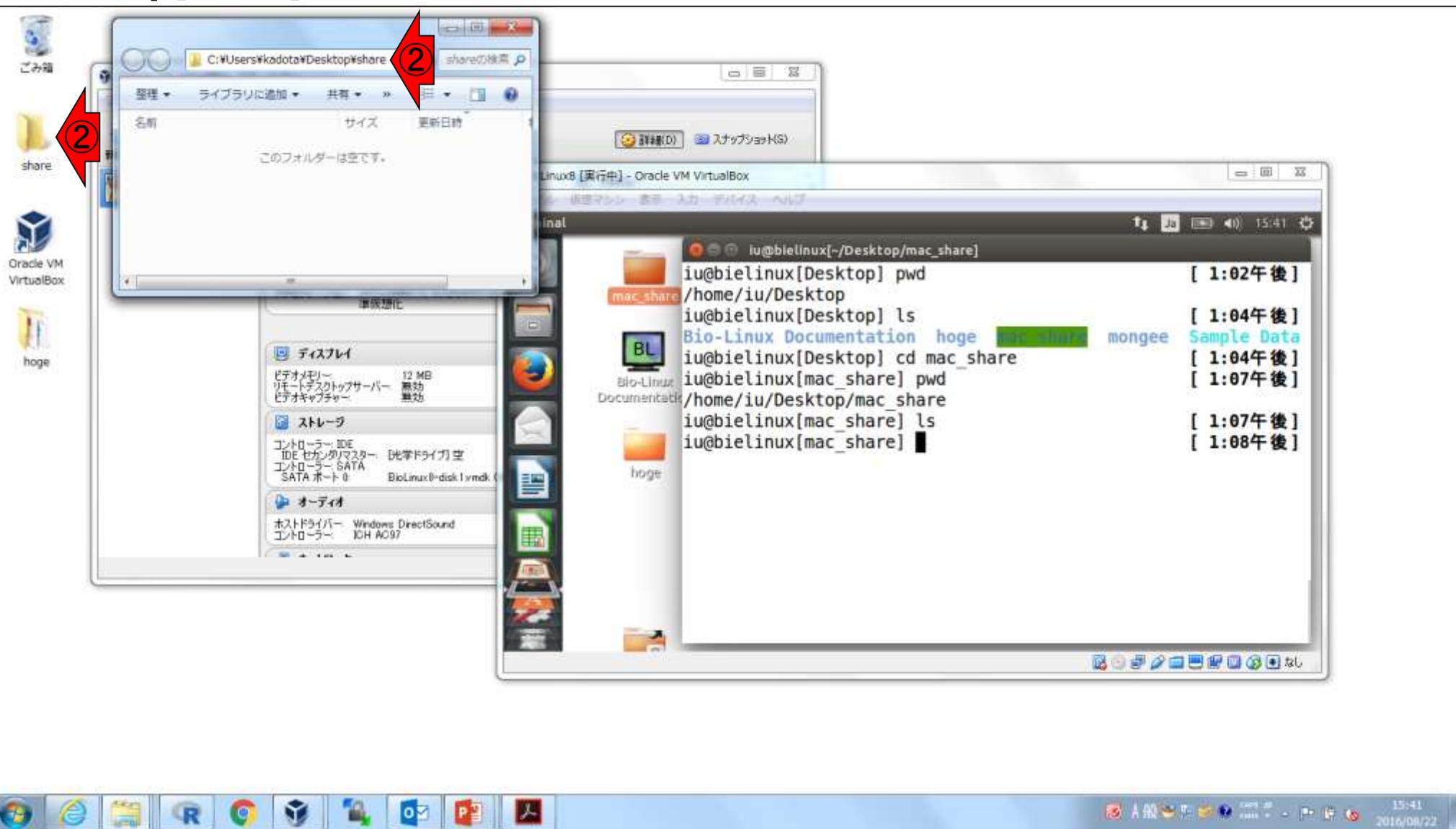
- グループごとに異なる課題ファイルを入力として、「ダウンロード、de novoアセンブリ、BLAST検索」を実行し、得られた結果をレポートにまとめて発表せよ
- グループ1はkadai1.fasta、グループ2はkadai2.fasta、etc.

# 共有フォルダ

①Linuxのmac\_shareと、②Windowsのshareは、共有フォルダです。③mac\_share上でlsした結果何もなかったなので、②には何もありません



# 共有フォルダ



# 共有フォルダ

①hogeフォルダ中の②pdfファイルを、③shareフォルダ内にコピーしてみましょう

The screenshot illustrates the process of copying files from a local folder to a shared folder. It shows three main components:

- File Explorer (share):** A window showing the 'share' folder, which is currently empty. A red arrow labeled '3' points to the empty space, indicating the destination for the copy.
- File Explorer (hoge):** A window showing the 'hoge' folder containing a list of PDF files. A red arrow labeled '2' points to the first file, 'JSLAB\_1\_kadota.pdf', indicating the source of the copy.
- Terminal:** A terminal window showing the execution of commands to copy files from the 'hoge' folder to the 'mac\_share' folder. The commands and their outputs are as follows:

```
iu@bielinux[~/Desktop/mac_share] pwd [ 1:02午後 ]
/home/iu/Desktop
iu@bielinux[~/Desktop] ls [ 1:04午後 ]
Bio-Linux Documentation hoge mongee Sample Data
iu@bielinux[~/Desktop] cd mac_share [ 1:04午後 ]
iu@bielinux[~/Desktop/mac_share] pwd [ 1:07午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[~/Desktop/mac_share] ls [ 1:07午後 ]
iu@bielinux[~/Desktop/mac_share] [ 1:08午後 ]
```

# 共有フォルダ

①こんな感じになります。共有フォルダなので、② mac\_share上でも同じファイルが見えるはず

The screenshot illustrates a file sharing setup. On the left, a Windows File Explorer window shows the 'share' folder containing 'JSLAB\_1\_kadota.pdf'. A red arrow labeled '1' points to this file. Below it, another File Explorer window shows the 'hoge' folder, which contains a 'style' subfolder and a list of PDF files, including 'JSLAB\_1\_kadota.pdf'. On the right, a Linux terminal window shows the following commands and output:

```
iu@bielinux[~/Desktop/mac_share]
iu@bielinux[Desktop] pwd
/home/iu/Desktop
iu@bielinux[Desktop] ls
Bio-Linux Documentation hoge mongee Sample Data
iu@bielinux[Desktop] cd mac_share
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls
iu@bielinux[mac_share]
```

A red arrow labeled '2' points to the terminal output, indicating that the same files are visible from the Linux side. The desktop taskbar at the bottom shows various application icons and the system tray with the date '2016/08/22' and time '15:46'.

# 共有フォルダ

②Isした結果、確かに見えました。こんな感じで、Linux上でのプログラム実行結果を共有フォルダ経由でWindowsに移動またはコピーし、Windows上で結果を整形するなどできます。共有フォルダについては、NGS連載第3-4回でも解説

The screenshot illustrates the process of accessing files from a Linux virtual machine through a shared folder on a Windows host. On the Windows side, the 'share' folder contains 'JSLAB\_1\_kadota.pdf'. The 'hoge' folder on the host contains a list of files including 'JSLAB\_1\_kadota.pdf' through 'JSLAB7\_suppl\_win\_20160620.pdf' and 'r\_seq.html'. The terminal window shows the user navigating to the shared folder 'mac\_share' and listing its contents, which matches the files in the 'hoge' folder.

```
iu@bielinux[~/Desktop/mac_share]
iu@bielinux[Desktop] pwd
/home/iu/Desktop
iu@bielinux[Desktop] ls
Bio-Linux Documentation hoge mongee Sample Data
iu@bielinux[Desktop] cd mac_share
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls
iu@bielinux[mac_share] ls
JSLAB_1_kadota.pdf
iu@bielinux[mac_share]
```

# Contents

## ■ イン트로ダクション

- 概要、背景(NGS用カリキュラム、講習会)、Linuxスキル習得の意義
- ウェブ情報(日本乳酸菌学会誌のNGS連載やNGS講習会資料)

## ■ 実習環境に慣れる

- 仮想環境での作業に慣れる
- GUIとCUI(マウス操作かコマンド入力操作か)
- ターミナルでの作業
- 共有フォルダの概念を理解

## ■ 練習

- 作業ディレクトリの変更、練習用NGSデータファイルのダウンロード
- ファイルの確認、de novoゲノムアセンブリ
- BLAST検索

## ■ 課題

- グループごとに異なる課題ファイルを入力として、「ダウンロード、de novoアセンブリ、BLAST検索」を実行し、得られた結果をレポートにまとめて発表せよ
- グループ1はkadai1.fasta、グループ2はkadai2.fasta、etc.

# 練習

練習用として①仮想NGSデータファイル(hoge.fasta)の、②ダウンロードから③de novoアセンブリ、および④BLAST検索から、アセンブリ結果として得られた塩基配列が乳酸菌ゲノム配列であることの確認までを行います

・ 門田幸二「講義資料」, 東京大学農学部展開科目: バイオインフォマテ

内容: バイオインフォマティクスやNGSをキーワード程度は知っているバイオインフォマティクスとLinuxスキル習得の意義。バイオインフォマティクス分野で需要が多いのはNGSデータの解析 (JST-NBDCIによるアンケート結果)。バイオインフォマティクス人材育成のための講習会の一環として行われているNGS講習会資料を有効活用するために必要な基本スキルの習得。具体的には、日本乳酸菌学会誌のNGS連載内容を理解するための基礎として、Bio-Linuxの習得に慣れるのが最低限の到達目標。何かやったという達成感を味わってもらうため、NGSデータのde novoゲノムアセンブリを行い、リード長(塩基配列の長さ)よりも伸びたことを実感してもらう。また、アセンブリ結果の一部をBLAST検索し、入力データ(hoge.fasta; 約6.4MB)がどんな生物種であったかを調べるあたりまで。課題はkadai1.fasta, kadai2.fasta, kadai3.fasta, kadai4.fastaのいずれかを選択して実行。下記と同様の手順で「データのダウンロード → de novoアセンブリ → BLAST検索」を行い、得られた結果を発表。約2日分。

#作業ディレクトリの変更

```
cd /home/iu/Desktop/mac_share
```

#解析したいファイル(hoge.fasta)のダウンロード

```
wget -c http://www.iu.a.u-tokyo.ac.jp/%7Ekadota/20160912/hoge.fasta
```

#行数やファイルサイズを表示(行数が200,000行、サイズが6,688,895 bytesとなっていればOK)

```
wc hoge.fasta
```

#最初の10行分を表示(2行分で1つのリードを表し、1リードが50塩基であることを確認)

```
head hoge.fasta
```

#de novoアセンブリを実行し、結果をugeディレクトリに保存

```
velveth uge 31 -short -fasta hoge.fasta
```

```
velvetg uge
```

#ugeディレクトリに移動して、アセンブリ結果ファイル(contigs.fa)があることを確認

```
cd uge
```

```
ls
```

#行数やファイルサイズを表示(入力のhoge.fastaは200,000行だったが、出力のcontigs.faは4,038行となっている)

```
wc contigs.fa
```

# 作業ディレクトリ

手順通りにやったヒトは、作業ディレクトリがmac\_shareのままであり、pdfファイルが1つある状態。この場合、①をやる必要はないが、やってもよいのでやってみる

・ 門田幸二、「講義資料」、東京大学農学部展開科目: バイオインフォマティクス, 東京大学(東京), 2016

内容: バイオインフォマティクスやNGSをキーワード程度は知っているヒト(学部3年生程度)向けのバイオインフォマティクスとLinuxスキル習得の意義。バイオインフォマティクス分野で需要が多いのはNGSデータの解析(JST-NBDCIによるアンケート結果)。バイオインフォマティクス人材育成のための講習会の一環として行われているNGS講習会資料を有効活用するために必要な基本スキルの習得。具体的には、日本乳酸菌学会誌のNGS連載内容を理解するための基礎として、Bio-Linuxのノリに慣れるのが最低限の到達目標。何かやったという達成感を味わってもらうため、NGSデータのde novoゲノムアセンブリを行い、元のリード長(塩基配列の長さ)よりも伸びたことを実感してもらう。また、アセンブリ結果の一部をBLAST検索し、入力データ(hoge.fasta: 約6.4MB)がどんな生物種であったかを調べるあたりまで。課題はkadai1.fasta, kadai2.fasta, kadai3.fasta, kadai4.fastaのいずれかを選択して実行。下記と同様の手順で「データのダウンロード → de novoアセンブリ → BLAST検索」し、得られた結果を発表。約2日分。

#作業ディレクトリの変更

cd /home/iu/Desktop/mac\_share

#解析したいファイル(hoge.fasta)のダウンロード

wget -c http://www.iu.a.u-tokyo.ac.jp/...

#行数やファイルサイズを表示(行数が200,000)

wc hoge.fasta

#最初の10行分を表示(2行分で1つのリードを1行)

head hoge.fasta

#de novoアセンブリを実行し、結果をugeディレクトリに保存

velveth uge 31 -short -fasta hoge.fasta

velvetg uge

#ugeディレクトリに移動して、アセンブリ結果を確認

cd uge

ls

#行数やファイルサイズを表示(入力のhoge.fasta)

wc contigs.fasta



```
File Edit View Search Terminal Help
iu@bielinux[mac_share] pwd [ 1:50午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls [ 1:50午後 ]
JSLAB_1_kadota.pdf
iu@bielinux[mac_share] [ 1:52午後 ]
```





# wgetでダウンロード

次は、①wgetというコマンドを用いて任意のURL上にあるファイル(hoge.fasta)のダウンロードです。②赤下線部分を丸々コピーでもいいのですが、せっかくなのでより汎用性の高い方法を伝授

・門田幸二「講義資料」, 東京大学農学部展開科目: バイオインフォマティクス, 東京大学

内容: バイオインフォマティクスやNGSをキーワード程度は知っているヒト(学部3年生)を対象とした、バイオインフォマティクスとLinuxスキル習得の意義。バイオインフォマティクス分野で需要が多いのはNGSデータの解析(JST-NBDCIによるアンケート結果)。バイオインフォマティクス人材育成のための講習会の一環として行われているNGS講習会資料を有効活用するために必要な基本スキルの習得。具体的には、日本乳酸菌学会誌のNGS連載内容を理解するための基礎として、Bio-Linuxのノリに慣れるのが最低限の到達目標。何かやったという達成感を味わってもらうため、NGSデータのde novoゲノムアセンブリを行い、元のリード長(塩基配列の長さ)よりも伸びたことを実感してもらう。また、アセンブリ結果の一部をBLAST検索し、入力データ(hoge.fasta; 約6.4MB)がどんな生物種であったかを調べるあたりまで。課題はkadai1.fasta, kadai2.fasta, kadai3.fasta, kadai4.fastaのいずれかを選択して実行。下記と同様の手順で「データのダウンロード → de novoアセンブリ → BLAST検索」し、得られた結果を発表。約2日分。

#作業ディレクトリの変更

```
cd /home/iu/Desktop/mac_share
```

#解析したいファイル(hoge.fasta)のダウンロード

```
wget -c http://www.iu.a.u-tokyo.ac.jp/%7Ekadota/20160912/hoge.fasta
```

#行数やファイルサイズを表示(行数が200,000行、サイズが6,688,895 bytesと7②ければOK)

```
wc hoge.fasta
```

#最初の10行分を表示(2行分で1つのリードを表し、1リードが50塩基であることを確認)

```
head hoge.fasta
```

#de novoアセンブリを実行し、結果をugeディレクトリに保存

```
velveth uge 31 -short -fasta hoge.fasta
```

```
velvetg uge
```

#ugeディレクトリに移動して、アセンブリ結果ファイル(contigs.fa)があることを確認

```
cd uge
```

```
ls
```

#行数やファイルサイズを表示(入力のhoge.fastaは200,000行だったが、出力のcontigs.faは4,038行となっている)

```
wc contigs.fa
```

# wgetでダウンロード

①「wget -c」(ダブルゲット、スペース、ハイフン、スペース)まで打ってから、②ダウンロードしたいファイル(hoge.fasta)のURL情報を取得

・ 門田幸二、「講義資料」、東京大学農学部展開科目: バイオインフォマティクス、東京大学(東京)、2016.09.12-16

内容: バイオインフォマティクスやNGSをキーワード程度は知っているヒト(学部3年生程度)向けのアクティブ・ラーニング系講義。バイオインフォマティクスとLinuxスキル習得の意義。バイオインフォマティクス分野で需要が多いのはNGSデータの解析(JST-NBDCによるアンケート結果)。バイオインフォマティクス人材育成のための講習会の一環として行われているNGS講習会資料を有効活用するために必要な基本スキルの習得。具体的には、日本乳酸菌学会誌のNGS連載内容を理解するための基礎として、Bio-Linuxの使い慣れるのが最低限の到達目標。何かやったという達成感を味わってもらうため、NGSデータのde novoゲノムアセンブリを行い、リーダー長(塩基配列の長さ)よりも伸びたことを実感してもらう。また、アセンブリ結果の一部をBLAST検索し、入力データ(hoge.fasta, 6.4MB)がどんな生物種であったかを調べるあたりまで。課題はkadai1.fasta, kadai2.fasta, kadai3.fasta, kadai4.fastaのいずれかを選択して実行。下記と同様の手順で「データのダウンロード → de novoアセンブリ → BLAST検索」し、得られた結果を発表。約2日分。

#作業ディレクトリの変更

```
cd /home/iu/Desktop/mac_share
```

#解析したいファイル(hoge.fasta)のダウンロード

```
wget -c http://www.iu.a.u-tokyo.ac.jp/
```

#行数やファイルサイズを表示(行数が200,000)

```
wc hoge.fasta
```

#最初の10行分を表示(2行分で1つのリードを1行)

```
head hoge.fasta
```

#de novoアセンブリを実行し、結果をugeディレクトリに保存

```
velveth uge 31 -short -fasta hoge.fasta  
velvetg uge
```

#ugeディレクトリに移動して、アセンブリ結果を確認

```
cd uge  
ls
```

#行数やファイルサイズを表示(入力のhoge.fasta)

```
wc contigs.fasta
```

```
File Edit View Search Terminal Help
iu@bielinux[mac_share] pwd [ 1:50午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls [ 1:50午後 ]
JSLAB_1_kadota.pdf
iu@bielinux[mac_share] cd /home/iu/Desktop/mac_share [ 2:04午後 ]
iu@bielinux[mac_share] pwd [ 2:04午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls [ 2:04午後 ]
JSLAB_1_kadota.pdf
iu@bielinux[mac_share] wget -c [ 2:04午後 ]
```

# wgetでダウンロード

①ダウンロードしたいファイル(hoge.fasta)上で右クリックして、②「ショートカットのコピー」。Macintoshの場合は「リンク先のコピー」だと思います

・ 門田幸二「講義資料」, 東京大学農学部展開科目: バイオインフォマティクス, 東京大学(東京), 2016.09.12-16

内容: バイオインフォマティクスやNGSをキーワード程度は知っているヒト(学部3年生程度)向けのアクティブ・ラーニング系講義。バイオインフォマティクスとLinuxスキル習得の意義。バイオインフォマティクス分野で需要が多いのはNGSデータの解析 (JST-NBDCによるアンケート結果)。バイオインフォマティクス人材育成のための講習会の一環として行われているNGS講習会資料を有効活用するために必要な基本スキルの習得。具体的には、日本乳酸菌学会誌のNGS連載内容を理解するための基礎として、Bio-Linuxのインストールに慣れるのが最低限の到達目標。何かやったという達成感を味わってもらうため、NGSデータのde novoゲノムアセンブリを行い、ダウンロード長(塩基配列の長さ)よりも伸びたことを実感してもらう。また、アセンブリ結果の一部をBLAST検索し、入力データ(hoge.fasta, 6.4MB)がどんな生物種であったかを調べるあたりまで。課題はkadai1.fasta, kadai2.fasta, kadai3.fasta, kadai4.fastaのいずれかを使用して実行。下記と同様の手順で「データのダウンロード → de novoアセンブリ → BLAST検索」し、得られた結果を発表。

#作業ディレクトリの変更

```
cd /home/iu/Desktop/mac_share
```

#解析したいファイル(hoge.fasta)のダウンロード

```
wget -c http://www.iu.a.u-tokyo.ac.jp/...
```

#行数やファイルサイズを表示(行数が200,000)

```
wc hoge.fasta
```

#最初の10行分を表示(2行分で1つのリードを1行)

```
head hoge.fasta
```

#de novoアセンブリを実行し、結果をugeディレクトリに保存

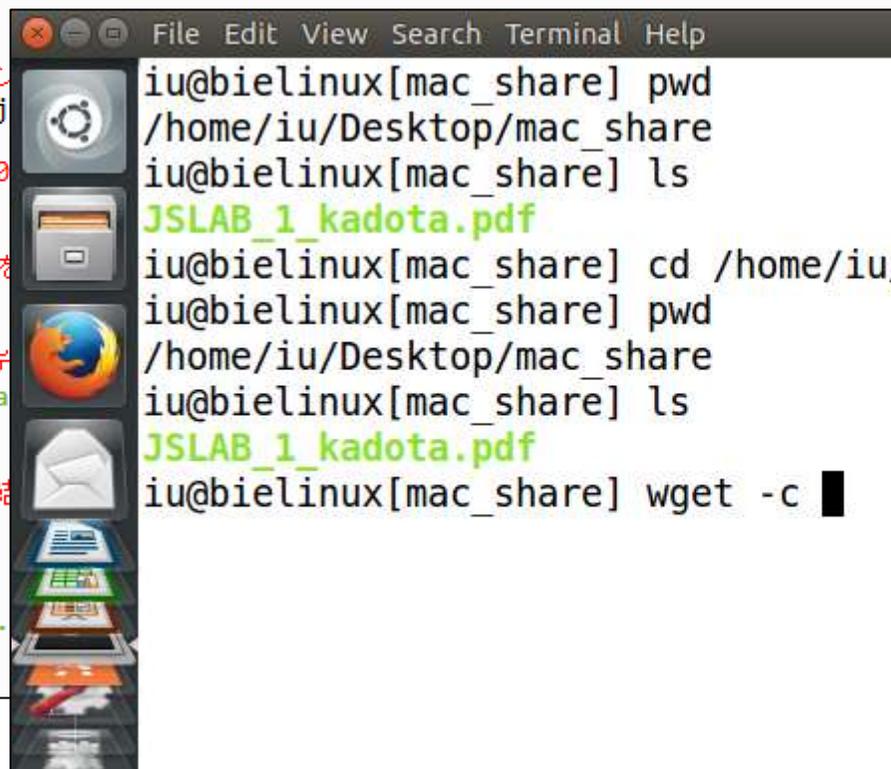
```
velveth uge 31 -short -fasta hoge.fasta  
velvetg uge
```

#ugeディレクトリに移動して、アセンブリ結果を表示

```
cd uge  
ls
```

#行数やファイルサイズを表示(入力のhoge.fasta)

```
wc contigs.fasta
```



```
File Edit View Search Terminal Help  
iu@bielinux[mac_share] pwd  
/home/iu/Desktop/mac_share  
iu@bielinux[mac_share] ls  
JSLAB_1_kadota.pdf  
iu@bielinux[mac_share] cd /home/iu/Desktop/mac_share  
iu@bielinux[mac_share] pwd  
/home/iu/Desktop/mac_share  
iu@bielinux[mac_share] ls  
JSLAB_1_kadota.pdf  
iu@bielinux[mac_share] wget -c
```



開く(O)  
新しいタブで開く(W)  
新しいウィンドウで開く(N)  
対象をファイルに保存(A)...  
対象を印刷(P)

切り取り  
コピー(C) ← ②  
ショートカットのコピー(T) ← ②  
貼り付け(P)

Bing で翻訳  
電子メール (Windows Live Hotmail)  
すべてのアクセラレータ

要素の検査(L)  
お気に入りに追加(F)...  
Send to OneNote  
プロパティ(R)

# wgetでダウンロード

・ 門田幸二, 「講義資料」, 東京大学農学部展開科目: バイオインフォマティクス, 東京大学(東京), 2016.09.12-16

内容: バイオインフォマティクスやNGSをキーワード程度は知っているヒト(学部3年生程度)向けのアクティブ・ラーニング系講義。バイオインフォマティクスとLinuxスキル習得の意義。バイオインフォマティクス分野で需要が多いのはNGSデータの解析 (JST-NBDCによるアンケート結果)。バイオインフォマティクス人材育成のための講習会の一環として行われているNGS講習会資料を有効活用するために必要な基本スキルの習得。具体的には、日本乳酸菌学会誌のNGS連載内容を理解するための基礎として、Bio-Linuxのノリに慣れるのが最低限の到達目標。何かやったという達成感を味わってもらうため、NGSデータのde novoゲノムアセンブリを行い、元のリード長(塩基配列の長さ)よりも伸びたことを実感してもらう。また、アセンブリ結果の一部をBLAST検索し、入力データ(hoge.fasta; 約6.4MB)がどんな生物種であったかを調べるあたりまで。課題はkadai1.fasta, kadai2.fasta, kadai3.fasta, kadai4.fastaのいずれかを選択して実行。下記と同様の手順で「データのダウンロード → de novoアセンブリ → BLAST検索」し、得られた結果を発表。約2日分。

#作業ディレクトリの変更

```
cd /home/iu/Desktop/mac_share
```

#解析したいファイル(hoge.fasta)のダウンロード

```
wget -c http://www.iu.a.u-tokyo.ac.jp/
```

#行数やファイルサイズを表示(行数が200,000)

```
wc hoge.fasta
```

#最初の10行分を表示(2行分で1つのリードを1行)

```
head hoge.fasta
```

#de novoアセンブリを実行し、結果をugeディレクトリに保存

```
velveth uge 31 -short -fasta hoge.fasta
velvetg uge
```

#ugeディレクトリに移動して、アセンブリ結果を確認

```
cd uge
ls
```

#行数やファイルサイズを表示(入力のhoge.fasta)

```
wc contigs.fasta
```

```
iu@bielinux[~/Desktop/mac_share]
iu@bielinux[mac_share] pwd [ 1:50午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls [ 1:50午後 ]
JSLAB_1_kadota.pdf
iu@bielinux[mac_share] cd /home/iu/Desktop/mac_share [ 2:04午後 ]
iu@bielinux[mac_share] pwd [ 2:04午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls [ 2:04午後 ]
JSLAB_1_kadota.pdf
iu@bielinux[mac_share] wget -c [ 2:04午後 ]
```

# wgetでダウンロード

①赤下線部分と同じURL情報を、②ペーストできていることがわかります。リターンキーを押すとダウンロードが始まります

・ 門田幸二、「講義資料」、東京大学農学部展開科目: バイオインフォマティクス, 東京大学(東京), 2016.09.12-16

内容: バイオインフォマティクスやNGSをキーワード程度は知っているヒト(学部3年生程度)向けのアクティブ・ラーニング系講義。バイオインフォマティクスとLinuxスキル習得の意義。バイオインフォマティクス分野で需要が多いのはNGSデータの解析 (JST-NBDCによるアンケート結果)。バイオインフォマティクス人材育成のための講習会の一環として行われているNGS講習会資料を有効活用するために必要な基本スキルの習得。具体的には、日本乳酸菌学会誌のNGS連載内容を理解するための基礎として、Bio-Linuxのノリに慣れるのが最低限の到達目標。何かやったという達成感を味わってもらうため、NGSデータのde novoゲノムアセンブリを行い、元のリード長(塩基配列の長さ)よりも伸びたことを実感してもらう。また、アセンブリ結果の一部をBLAST検索し、入力データ(hoge.fasta; 約6.4MB)がどんな生物種であったかを調べるあたりまで。課題はkadai1.fasta, kadai2.fasta, kadai3.fasta, kadai4.fastaのいずれかを選択して実行。下記と同様の手順で「データのダウンロード → de novoアセンブリ → BLAST検索」し、得られた結果を発表。約2日分。

```
#作業ディレクトリの変更
cd /home/iu/Desktop/mac_share

#解析したいファイル(hoge.fasta)のダウンロード
wget -c http://www.iu.a.u-tokyo.ac.jp/~kadota/20160912/hoge.fasta

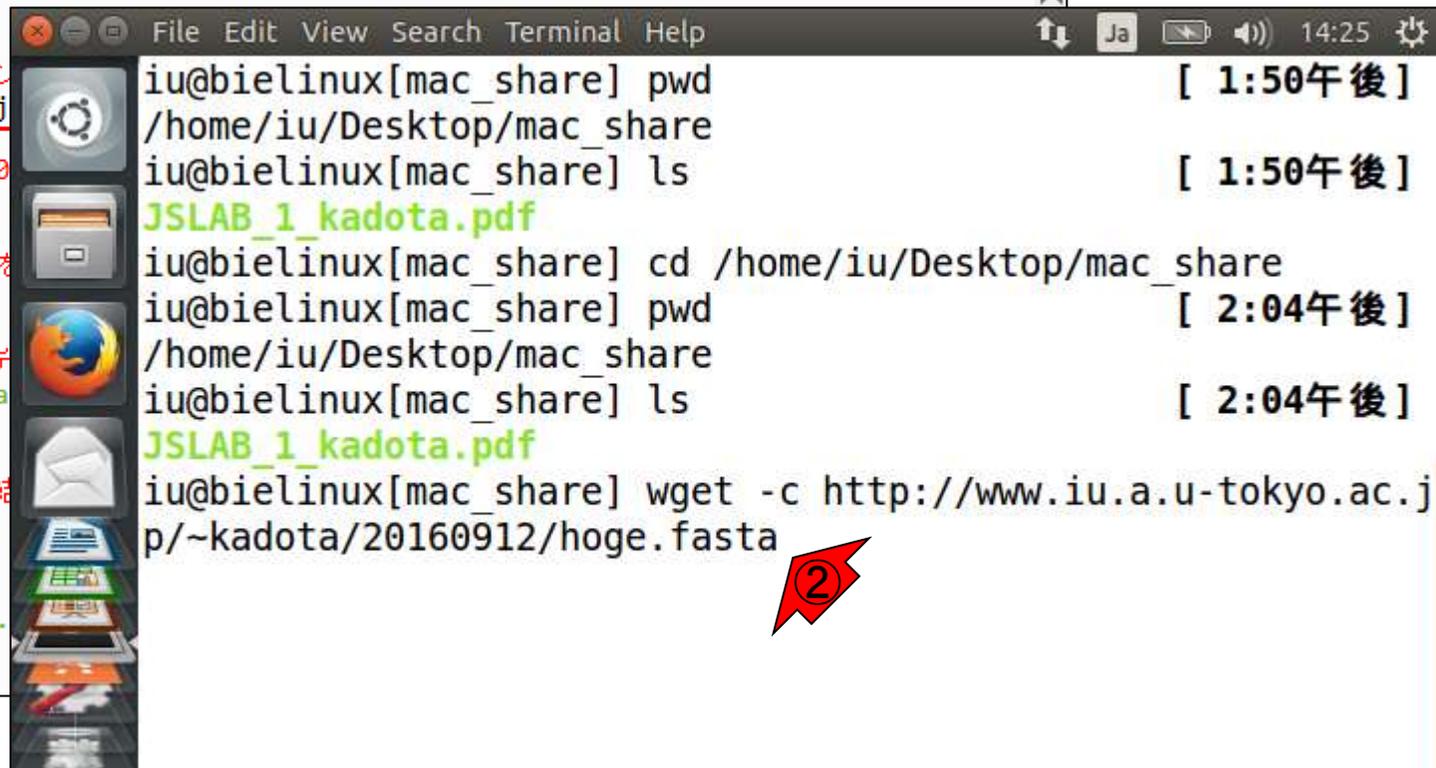
#行数①とファイルサイズを表示(行数が200,000)
wc hoge.fasta

#最初の10行分を表示(2行分で1つのリードを表示)
head hoge.fasta

#de novoアセンブリを実行し、結果をugeディレクトリに保存
velveth uge 31 -short -fasta hoge.fasta
velvetg uge

#ugeディレクトリに移動して、アセンブリ結果を表示
cd uge
ls

#行数やファイルサイズを表示(入力のhoge.fasta)
wc contigs.fasta
```



```
File Edit View Search Terminal Help
iu@bielinux[mac_share] pwd [ 1:50午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls [ 1:50午後 ]
JSLAB_1_kadota.pdf
iu@bielinux[mac_share] cd /home/iu/Desktop/mac_share
iu@bielinux[mac_share] pwd [ 2:04午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls [ 2:04午後 ]
JSLAB_1_kadota.pdf
iu@bielinux[mac_share] wget -c http://www.iu.a.u-tokyo.ac.jp/~kadota/20160912/hoge.fasta
```

# wget実行直後

- 門田幸二「講義資料」, 東京大学農学部展開科目: バイオインフォマティクス, 東京大学(東京), 2016.09.12-16
- 内容: バイオインフォマティクスやNGSをキーワード程度は知っているヒト(学部3年生程度)向けのアクティブ・ラーニング系講義。バイオインフォマティクスとLinuxスキル習得の意義。バイオインフォマティクス分野で需要が多いのはNGSデータの解析 (JST-NBDCによるアンケート結果)。バイオインフォマティクス人材育成のための講習会の一環として行われているNGS講習会資料を有効活用するために必要な基本スキルの習得。具体的には、日本乳酸菌学会誌のNGS連載内容を理解するための基礎として、Bio-Linuxのノリに慣れるのが最低限の到達目標。何かやったという達成感を味わってもらうため、NGSデータのde novoゲノムアセンブリを行い、元のリード長(塩基配列の長さ)よりも伸びたことを実感してもらう。また、アセンブリ結果の一部をBLAST検索し、入力データ(hoge.fasta; 約6.4MB)がどんな生物種であったかを調べるあたりまで。課題はkadai1.fasta, kadai2.fasta, kadai3.fasta, kadai4.fastaのいずれかを選択して実行。下記と同様の手順で「データのダウンロード → de novoアセンブリ → BLAST検索」し、得られた結果を発表。約2日分。

```
#作業ディレクトリの変更
cd /home/iu/Desktop/mac_share

#解析したいファイル(hoge.fasta)のダウンロード
wget -c http://www.iu.a.u-tokyo.ac.jp/ota/20160912/hoge.fasta

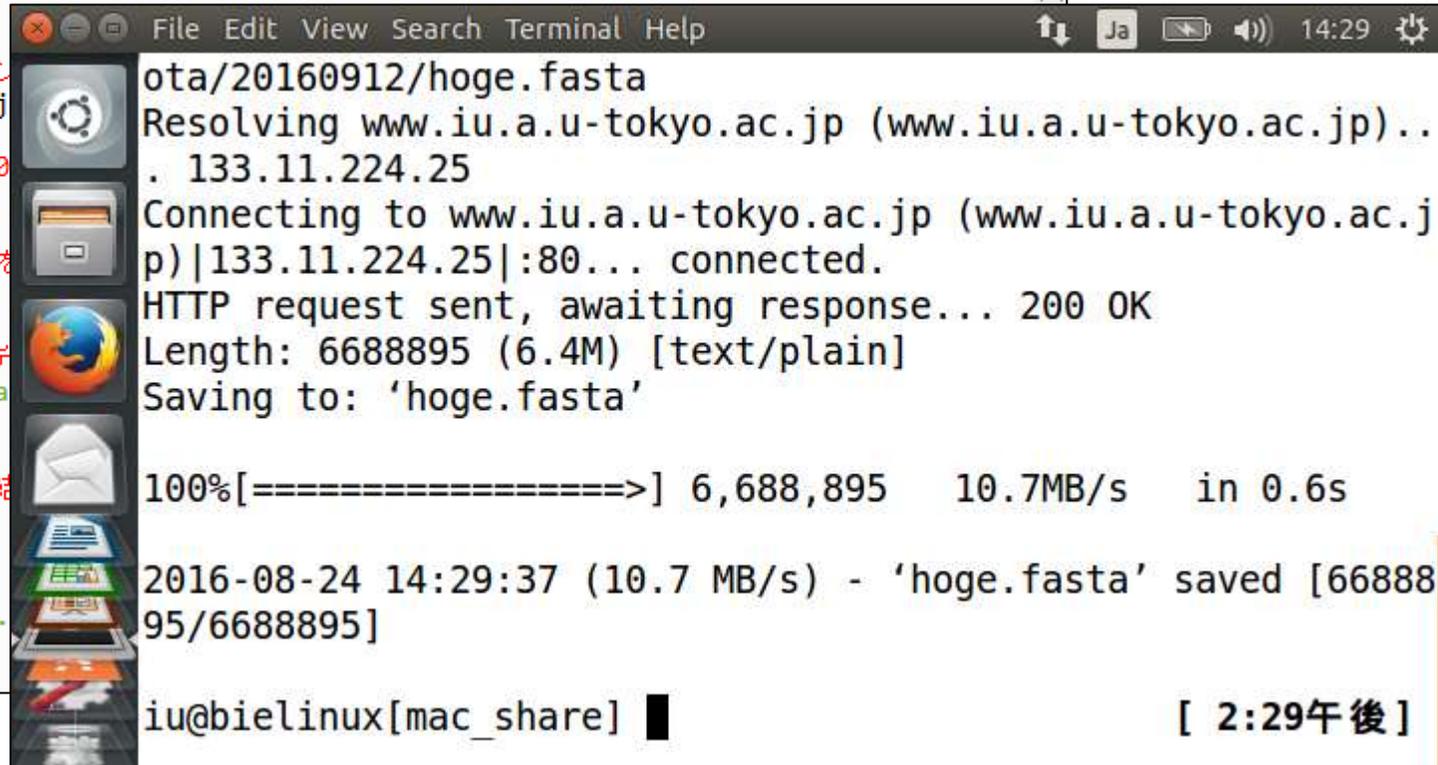
#行数やファイルサイズを表示(行数が200,000)
wc hoge.fasta

#最初の10行分を表示(2行分で1つのリードを表示)
head hoge.fasta

#de novoアセンブリを実行し、結果をugeディレクトリに保存
velveth uge 31 -short -fasta hoge.fasta
velvetg uge

#ugeディレクトリに移動して、アセンブリ結果を確認
cd uge
ls

#行数やファイルサイズを表示(入力のhoge.fasta)
wc contigs.fasta
```



```
File Edit View Search Terminal Help
ota/20160912/hoge.fasta
Resolving www.iu.a.u-tokyo.ac.jp (www.iu.a.u-tokyo.ac.jp)..
. 133.11.224.25
Connecting to www.iu.a.u-tokyo.ac.jp (www.iu.a.u-tokyo.ac.jp)|133.11.224.25|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 6688895 (6.4M) [text/plain]
Saving to: 'hoge.fasta'

100%[=====>] 6,688,895 10.7MB/s in 0.6s

2016-08-24 14:29:37 (10.7 MB/s) - 'hoge.fasta' saved [6688895/6688895]

iu@bielinux[mac_share] [ 2:29午後 ]
```

# Contents

## ■ イン트로ダクション

- 概要、背景(NGS用カリキュラム、講習会)、Linuxスキル習得の意義
- ウェブ情報(日本乳酸菌学会誌のNGS連載やNGS講習会資料)

## ■ 実習環境に慣れる

- 仮想環境での作業に慣れる
- GUIとCUI(マウス操作かコマンド入力操作か)
- ターミナルでの作業
- 共有フォルダの概念を理解

## ■ 練習

- 作業ディレクトリの変更、練習用NGSデータファイルのダウンロード
- ファイルの確認、de novoゲノムアセンブリ
- BLAST検索

## ■ 課題

- グループごとに異なる課題ファイルを入力として、「ダウンロード、de novoアセンブリ、BLAST検索」を実行し、得られた結果をレポートにまとめて発表せよ
- グループ1はkadai1.fasta、グループ2はkadai2.fasta、etc.

# lsで確認

①ls(えるえす)で確認。②確かにダウンロードしたhoge.fastaがあります。③ls-l(エルएस、スペース、ハイフンえる)でより詳細な情報を見ることができます

・門田幸二、「講義資料」, 東京大学農学部展開科目: バイオインフォマティクス, 東京大学(東京), 2016  
内容: バイオインフォマティクスやNGSをキーワード程度は知っているヒト(学部3年生程度)向けのバイオインフォマティクスとLinuxスキル習得の意義。バイオインフォマティクス分野で需要が多いのはNGSデータの解析(JST-NBDCIによるアンケート結果)。バイオインフォマティクス人材育成のための講習会の一環として行われているNGS講習会資料を有効活用するために必要な基本スキルの習得。具体的には、日本乳酸菌学会誌のNGS連載内容を理解するための基礎として、Bio-Linuxのノリに慣れるのが最低限の到達目標。何かやったという達成感を味わってもらうため、NGSデータのde novoゲノムアセンブリを行い、元のリード長(塩基配列の長さ)よりも伸びたことを実感してもらう。また、アセンブリ結果の一部をBLAST検索し、入力データ(hoge.fasta; 約6.4MB)がどんな生物種であったかを調べるあたりまで。課題はkadai1.fasta, kadai2.fasta, kadai3.fasta, kadai4.fastaのいずれかを選択して実行。下記と同様の手順で「データのダウンロード → de novoアセンブリ → BLAST検索」し、得られた結果を発表。約2日分。

```
#作業ディレクトリの変更
cd /home/iu/Desktop/mac_share

#解析したいファイル(hoge.fasta)のダウンロード
wget -c http://www.iu.a.u-tokyo.ac.jp/...

#行数やファイルサイズを表示(行数が200,000)
wc hoge.fasta

#最初の10行分を表示(2行分で1つのリードを表示)
head hoge.fasta

#de novoアセンブリを実行し、結果をugeディレクトリに保存
velveth uge 31 -short -fasta hoge.fasta -unlabeled-min 1000000000
velvetg uge

#ugeディレクトリに移動して、アセンブリ結果を確認
cd uge
ls

#行数やファイルサイズを表示(入力されたhoge.fasta)
wc contigs.fasta
```

```
File Edit View Search Terminal Help
Length: 6688895 (6.4M) [text/plain]
Saving to: 'hoge.fasta'

100%[=====>] 6,688,895 10.7MB/s in 0.6s

2016-08-24 14:29:37 (10.7 MB/s) - 'hoge.fasta' saved [6688895/6688895]

iu@bielinux[mac_share] ls [ 2:29午後 ]
hoge.fasta JSLAB_1_kadota.pdf
iu@bielinux[mac_share] ls -l [ 2:35午後 ]
total 7764
-rwxrwxrwx 1 iu iu 6688895 8月 22 18:19 hoge.fasta
-rwxrwxrwx 1 iu iu 1260591 8月 3 2014 JSLAB_1_kadota.pdf
iu@bielinux[mac_share] █ [ 2:35午後 ]
```

# WCで確認

①wcコマンドは、主にファイルの行数を調べる目的で利用します。②確かに200,000行になっていることがわかります。NGSデータの場合は、行数から「(リードと呼ばれる)塩基配列の数」を調べることができます

```
#作業ディレクトリの変更
cd /home/iu/Desktop/mac_share
```

```
#解析したいファイル(hoge.fasta)のダウンロード
wget -c http://www.iu.a.u-tokyo.ac.jp/%7Ekadota/20160912/hoge.fasta
```

```
#行数やファイルサイズを表示(行数が200,000行、サイズが6,688,895 bytesとなっていればOK)
wc hoge.fasta
```

```
#最初の10行分を表示(2行分で1つのリードを表し、1リードが50塩基であることを確認)
head hoge.fasta
```

```
#de novoアセンブリを実行し、結果をugeディレクトリに保存
velveth uge 31 -short -fasta hoge.fasta
velvetg uge
```

```
#ugeディレクトリに移動して、アセンブリ結果を確認
cd uge
ls
```

```
#行数やファイルサイズを表示(入力のhoge.fasta)
wc contigs.fasta
```

```
File Edit View Search Terminal Help
100%[=====>] 6,688,895 10.7MB/s in 0.6s
2016-08-24 14:29:37 (10.7 MB/s) - 'hoge.fasta' saved [6688895/6688895]
iu@bielinux[mac_share] ls [ 2:29午後 ]
hoge.fasta JSLAB_1_kadota.pdf
iu@bielinux[mac_share] ls -l [ 2:35午後 ]
total 7764
-rwxrwxrwx 1 iu iu 6688895 8月 22 18:19 hoge.fasta
-rwxrwxrwx 1 iu iu 1260591 8月 3 2014 JSLAB_1_kadota.pdf
iu@bielinux[mac_share] wc hoge.fasta [ 2:35午後 ]
200000 200000 6688895 hoge.fasta
iu@bielinux[mac_share]
```

# WCで確認

①ファイルサイズ情報。②ls -l実行結果として得られる、③の値と同じです

```
#作業ディレクトリの変更
cd /home/iu/Desktop/mac_share

#解析したいファイル(hoge.fasta)のダウンロード
wget -c http://www.iu.a.u-tokyo.ac.jp/%7Ekadota/20160912/hoge.fasta

#行数やファイルサイズを表示(行数が200,000行、サイズが6,688,895 bytesとなっていればOK)
wc hoge.fasta

#最初の10行分を表示(2行分で1つのリードを表し、1リードが50塩基であることを確認)
head hoge.fasta
```

```
#de novoアセンブリを実行し、結果をugeディレクトリに保存
velveth uge 31 -short -fasta hoge.fasta
velvetg uge

#ugeディレクトリに移動して、アセンブリ結果を確認
cd uge
ls

#行数やファイルサイズを表示(入力のhoge.fasta)
wc contigs.fasta
```

```
File Edit View Search Terminal Help
100%[=====>] 6,688,895 10.7MB/s in 0.6s
2016-08-24 14:29:37 (10.7 MB/s) - 'hoge.fasta' saved [6688895/6688895]
iu@bielinux[mac_share] ls [ 2:29午後 ]
hoge.fasta JSLAB_1_kadota.pdf
iu@bielinux[mac_share] ls -l [ 2:35午後 ]
total 7764
-rwxrwxrwx 1 iu iu 6688895 8月 22 18:19 hoge.fasta
-rwxrwxrwx 1 iu iu 1200000 8月 3 2014 JSLAB_1_kadota.pdf
iu@bielinux[mac_share] wc hoge.fasta [ 2:35午後 ]
200000 200000 6688895 hoge.fasta
iu@bielinux[mac_share] [ 2:45午後 ]
```

# headで確認

①headは、(デフォルトでは)ファイルの最初の10行分を表示させるコマンドです。このファイルは、FASTA形式と呼ばれるもので、2行で1つのリードを表します

```
#作業ディレクトリの変更  
cd /home/iu/Desktop/mac_share
```

```
#解析したいファイル(hoge.fasta)のダウンロード  
wget -c http://www.iu.a.u-tokyo.ac.jp/%7Ekadota/20160912/hoge.fasta
```

```
#行数やファイルサイズを表示(行数が200,000行、サイズが6,688,895 bytesとなっていればOK)  
wc hoge.fasta
```

```
#最初の10行分を表示(1行分で1つのリードを表し、1リードが50塩基であることを確認)  
head hoge.fasta
```

```
#de novoアセンブリを実行し、結果をugeディレクトリに保存  
velveth uge 31 -short -fasta hoge.fasta  
velvetg uge
```

```
#ugeディレクトリに移動して、アセンブリ結果を確認  
cd uge  
ls
```

```
#行数やファイルサイズを表示(入力のhoge.fasta)  
wc contigs.fasta
```

```
File Edit View Search Terminal Help 15:19
-rwxrwxrwx 1 iu iu 1260591  8月  3  2014 JSLAB_1_kadota.pdf
iu@bielinux[mac_share] wc hoge.fasta [ 2:35午後 ]
200000 200000 6688895 hoge.fasta
iu@bielinux[mac_share] head hoge.fasta [ 2:45午後 ]
>sequence.1
ATGNATCGAAACAGTATTTACAAGATTTGCATACTGAAATTGAAGCTGAT
>sequence.2
GTCNGAACACATGAATGGTGAACGGCGCTGAACTTTTCACGGACGCGGC
>sequence.3
CAANGATACAATCATTATCATGAACTCTAATGCCGGTTCTGGTGATGCAG
>sequence.4
AGANAGTATTTATCGTAAAATCTTAATTCATTCACAAGGTGGGGCGAACC
>sequence.5
AGTNCGGTTAGTGGAAGCTGCTAGGCATGTGCACACACCATTGTTTGCAC
iu@bielinux[mac_share] [ 3:19午後 ]
```

# headで確認

10塩基ごとに灰色の縦線を入れています。このNGSデータは、(少なくともここで見えている最初の5リード分については)50塩基の長さであることがわかります。①このファイル(hoge.fasta)は、②200,000行からなるので100,000リード。このように、大量の短いリード(short read)からなるのが典型的なNGSデータ

```
#作業ディレクトリの変更
cd /home/iu/Desktop/mac_share
```

```
#解析したいファイル(hoge.fasta)のダウンロード
wget -c http://www.iu.a.u-tokyo.ac.jp/%7Ekadota/20160912/hoge.fasta
```

```
#行数やファイルサイズを表示(行数が200,000行、サイズが6,688,895 bytesとなっていればOK)
wc hoge.fasta
```

```
#最初の10行分を表示(2行分で1つのリードを表し、1リードが50塩基であることを確認)
head hoge.fasta
```

```
#de novoアセンブリを実行し、結果をugeディレクトリに保存
velveth uge 31 -short -fasta hoge.fasta
velvetg uge
```

```
#ugeディレクトリに移動して、アセンブリ結果を確認
cd uge
ls
```

```
#行数やファイルサイズを表示(入力のhoge.fasta)
wc contigs.fasta
```

```
File Edit View Search Terminal Help
-rwxrwxrwx 1 iu iu 1260591  8月  3  2014  JSLAB_1_kadota.pdf
iu@bielinux[mac_share] wc hoge.fasta           [ 2:35午後 ]
200000 200000 6688895 hoge.fasta
iu@bielinux[mac_share] head hoge.fasta         [ 2:45午後 ]
>sequence.1
ATGNATCGAAACAGTATTTACAAGATTTGCATACTGAAATTGAAGCTGAT
>sequence.2
GTCNGAACACATGAATGGTGAACGGCGCTGAACTTTTCACGGACGCGGC
>sequence.3
CAANGATACAATCATTATCATGAACTCTAATGCCGGTTCTGGTGATGCAG
>sequence.4
AGANAGTATTTATCGTAAAATCTTAATTCATTCACAAGGTGGGGCGAACC
>sequence.5
AGTNCGGTTAGTGGAAGCTGCTAGGCATGTGCACACACCATTGTTTGCAC
iu@bielinux[mac_share] [ 3:19午後 ]
```

# Contents

## ■ イントロダクション

- 概要、背景(NGS用カリキュラム、講習会)、Linuxスキル習得の意義
- ウェブ情報(日本乳酸菌学会誌のNGS連載やNGS講習会資料)

## ■ 実習環境に慣れる

- 仮想環境での作業に慣れる
- GUIとCUI(マウス操作かコマンド入力操作か)
- ターミナルでの作業
- 共有フォルダの概念を理解

## ■ 練習

- 作業ディレクトリの変更、練習用NGSデータファイルのダウンロード
- ファイルの確認、de novoゲノムアセンブリ
- BLAST検索

## ■ 課題

- グループごとに異なる課題ファイルを入力として、「ダウンロード、de novoアセンブリ、BLAST検索」を実行し、得られた結果をレポートにまとめて発表せよ
- グループ1はkadai1.fasta、グループ2はkadai2.fasta、etc.

# de novoアセンブリ

①このデータは乳酸菌(*Lactobacillus hokkaidonensis*)ゲノムの実際のNGSデータの一部。NGSデータ解析の目的の1つは、このような短い塩基配列データを入力として、より長い元のゲノム配列を再構築すること

```
#作業ディレクトリの変更
cd /home/iu/Desktop/mac_share
```

```
#解析したいファイル(hoge.fasta)のダウンロード
wget -c http://www.iu.a.u-tokyo.ac.jp/%7Ekadota/20160912/hoge.fasta
```

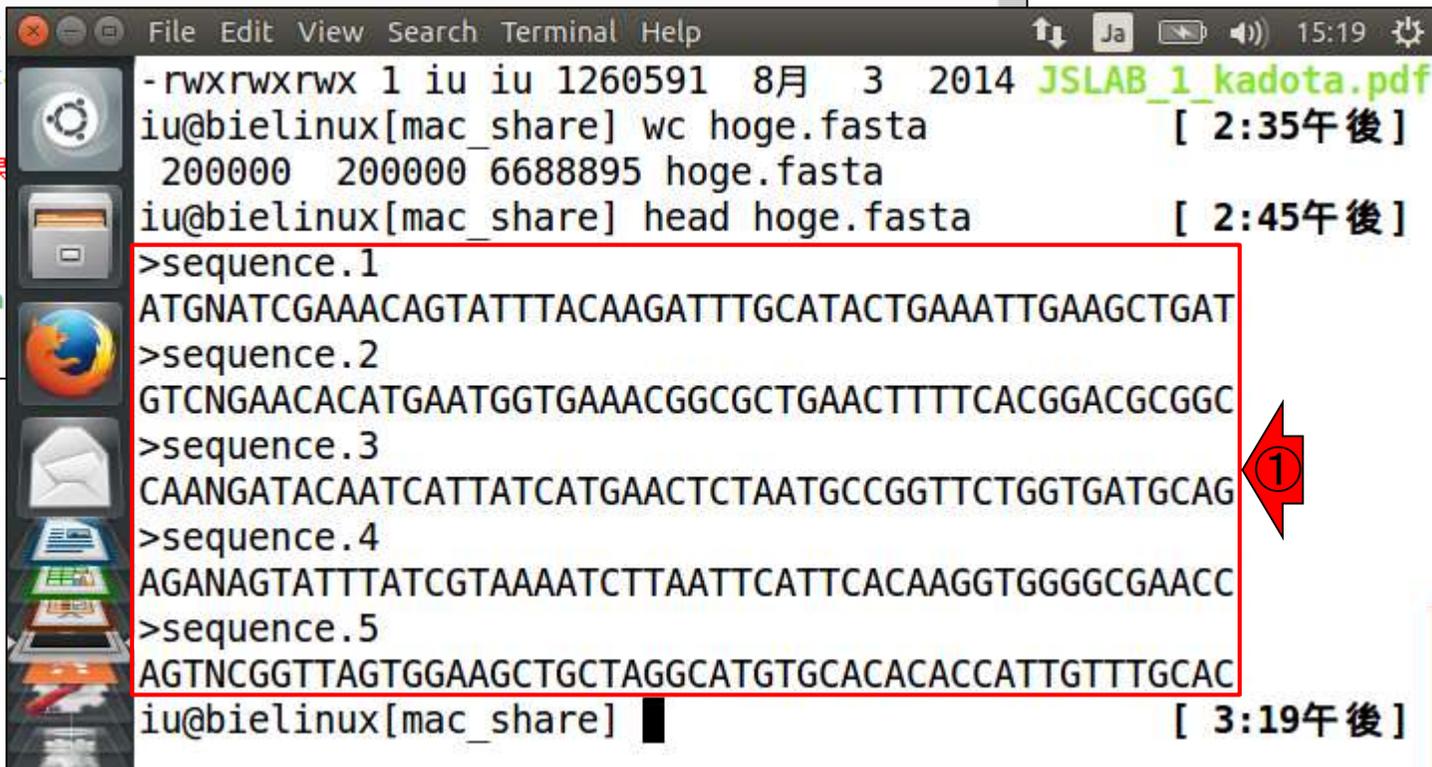
```
#行数やファイルサイズを表示(行数が200,000行、サイズが6,688,895 bytesとなっていればOK)
wc hoge.fasta
```

```
#最初の10行分を表示(2行分で1つのリードを表し、1リードが50塩基であることを確認)
head hoge.fasta
```

```
#de novoアセンブリを実行し、結果をugeディレクトリに保存
velveth uge 31 -short -fasta hoge.fasta
velvetg uge
```

```
#ugeディレクトリに移動して、アセンブリ結果を確認
cd uge
ls
```

```
#行数やファイルサイズを表示(入力のhoge.fastaと結果のcontigs.fasta)
wc contigs.fasta
```



```
File Edit View Search Terminal Help
-rwxrwxrwx 1 iu iu 1260591  8月  3  2014  JSLAB_1_kadota.pdf
iu@bielinux[mac_share] wc hoge.fasta           [ 2:35午後 ]
 200000  200000 6688895 hoge.fasta
iu@bielinux[mac_share] head hoge.fasta        [ 2:45午後 ]
>sequence.1
ATGNATCGAAACAGTATTTACAAGATTTGCATACTGAAATTGAAGCTGAT
>sequence.2
GTCNGAACACATGAATGGTCAAACGGCGCTGAACTTTTCACGGACGCGGC
>sequence.3
CAANGATACAATCATTATCATGAACTCTAATGCCGGTTCTGGTGATGCAG
>sequence.4
AGANAGTATTTATCGTAAAATCTTAATTCATTCACAAGGTGGGGCGAACC
>sequence.5
AGTNCGGTTAGTGGAAGCTGCTAGGCATGTGCACACACCATTGTTTGCAC
iu@bielinux[mac_share] [ 3:19午後 ]
```

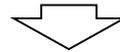
# de novoアセンブリ

入出力のイメージ。de novoアセンブリとは、リードの塩基配列情報のみを頼りに、元のリード長よりも長い配列(コンティグ)を出力する作業。この例の場合、赤下線が一致部分。出力は、元のリード長よりも2塩基長いコンティグとなる

入力: NGSリードファイル

リード1: CACCAGGACATGAAGACGCG

リード2: CCAGGACATGAAGACGCGTT



出力: コンティグ(より長くなった塩基配列)

CACCAGGACATGAAGACGCGTT

# de novoアセンブリ

①赤枠部分をコピー実行。Velvetというアセンブリプログラムを実行しているが、細かいコマンドの意味などはここでは気にしなくてよい。ここで重要なのは、入力はhoge.fastaであり、プログラムを実行するとugeというディレクトリが作成されるということのみ。そしてugeディレクトリ内にあるcontigs.faが主なアセンブリ結果ファイル

```
#作業ディレクトリの変更
cd /home/iu/Desktop/mac_share

#解析したいファイル(hoge.fasta)のダウンロード
wget -c http://www.iu.a.u-tokyo.ac.jp/%7Ekadota/20160912/
```

```
#行数やファイルサイズを表示(行数が200,000行、サイズが6,688,895 bytesとなっていればOK)
wc hoge.fasta
```

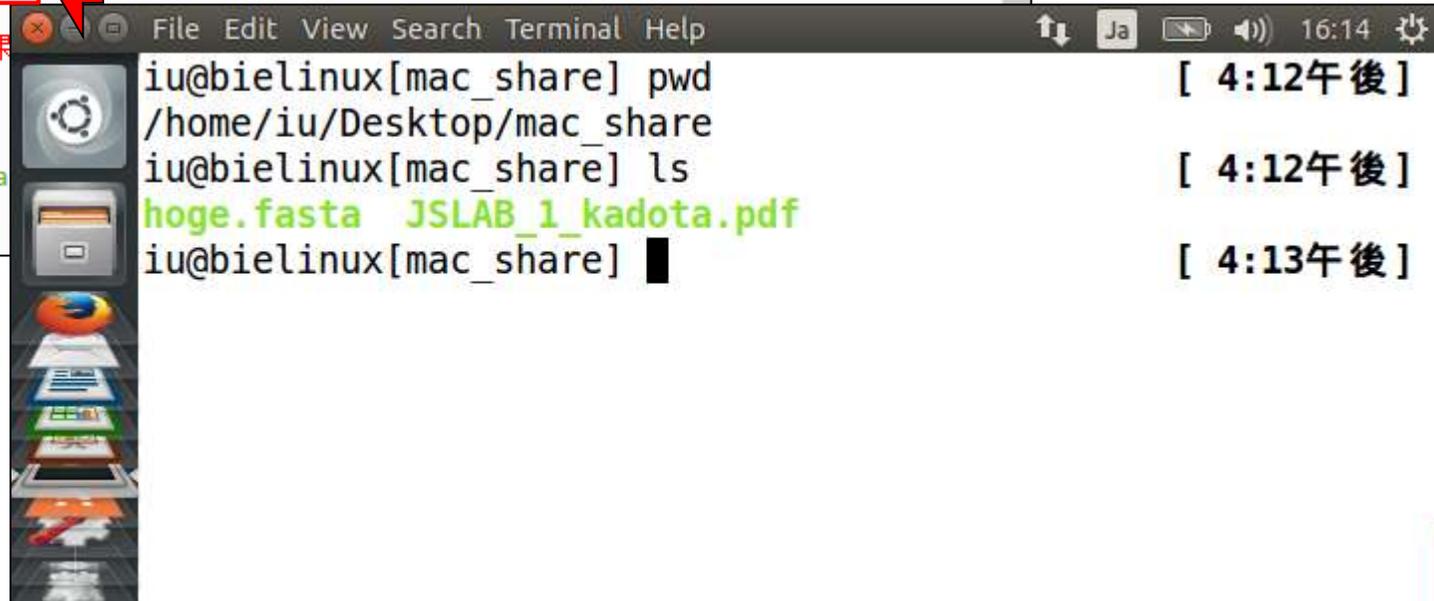
```
#最初の10行分を表示(2行分で1つのリードを表し、1リードが50塩基であることを確認)
head hoge.fasta
```

```
#de novoアセンブリを実行し、結果をugeディレクトリに保存
```

```
velveth uge 31 -short -fasta hoge.fasta
velvetg uge
```

```
#ugeディレクトリに移動して、アセンブリ結果を確認
cd uge
ls
```

```
#行数やファイルサイズを表示(入力のhoge.fasta)
wc contigs.fa
```



The screenshot shows a terminal window with the following content:

```
File Edit View Search Terminal Help
iu@bielinux[mac_share] pwd [ 4:12午後]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls [ 4:12午後]
hoge.fasta JSLAB_1_kadota.pdf
iu@bielinux[mac_share] [ 4:13午後]
```

A red circle with the number 1 is overlaid on the terminal output, pointing to the `velveth uge 31 -short -fasta hoge.fasta` command.

①コピー実行後の状態。計算自体は10秒程度で終わります

# コピー実行直後

```
#作業ディレクトリの変更
cd /home/iu/Desktop/mac_share

#解析したいファイル(hoge.fasta)のダウンロード
wget -c http://www.iu.a.u-tokyo.ac.jp/%7Ekadota/20160912/hoge.fasta

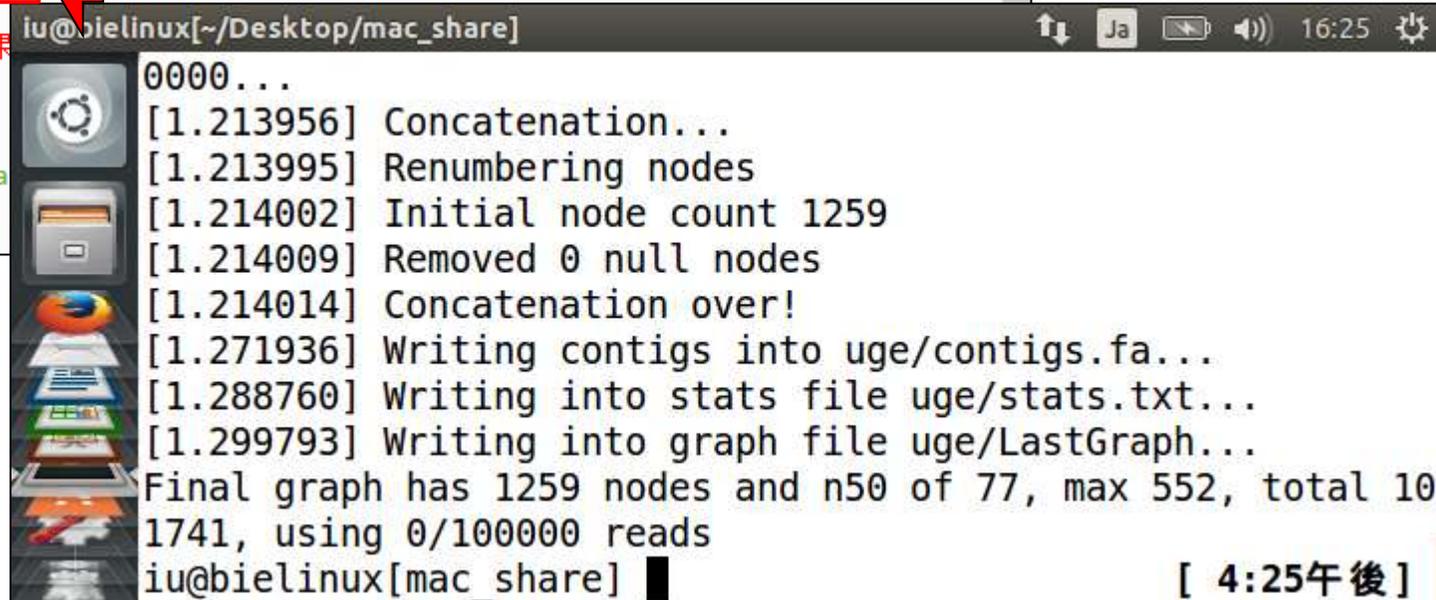
#行数やファイルサイズを表示(行数が200,000行、サイズが6,688,895 bytesとなっていればOK)
wc hoge.fasta

#最初の10行分を表示(2行分で1つのリードを表し、1リードが50塩基であることを確認)
head hoge.fasta

#de novoアセンブリを実行し、結果をugeディレクトリに保存
velveth uge 31 -short -fasta hoge.fasta
velvetg uge
```

```
#ugeディレクトリに移動して、アセンブリ結果
cd uge
ls

#行数やファイルサイズを表示(入力のhoge.fasta)
wc contigs.fa
```



```
iu@bielinux[~/Desktop/mac_share]
0000...
[1.213956] Concatenation...
[1.213995] Renumbering nodes
[1.214002] Initial node count 1259
[1.214009] Removed 0 null nodes
[1.214014] Concatenation over!
[1.271936] Writing contigs into uge/contigs.fa...
[1.288760] Writing into stats file uge/stats.txt...
[1.299793] Writing into graph file uge/LastGraph...
Final graph has 1259 nodes and n50 of 77, max 552, total 10
1741, using 0/100000 reads
iu@bielinux[mac_share]
```

## lsで確認

```
#作業ディレクトリの変更
cd /home/iu/Desktop/mac_share

#解析したいファイル(hoge.fasta)のダウンロード
wget -c http://www.iu.a.u-tokyo.ac.jp/%7Ekadota/20160912/hoge.fasta

#行数やファイルサイズを表示(行数が200,000行、サイズが6,688,895 bytesとなっていればOK)
wc hoge.fasta

#最初の10行分を表示(2行分で1つのリードを表し、1リードが50塩基であることを確認)
head hoge.fasta

#de novoアセンブリを実行し、結果をugeディレクトリに保存
velveth uge 31 -short -fasta hoge.fasta
velvetg uge

#ugeディレクトリに移動して、アセンブリ結果を確認
cd uge
ls

#行数やファイルサイズを表示(入力のhoge.fasta)
wc contigs.fa
```

```
File Edit View Search Terminal Help
[1.213995] Renumbering nodes
[1.214002] Initial node count 1259
[1.214009] Removed 0 null nodes
[1.214014] Concatenation over!
[1.271936] Writing contigs into uge/contigs.fa...
[1.288760] Writing into stats file uge/stats.txt...
[1.299793] Writing into graph file uge/LastGraph...
Final graph has 1259 nodes and n50 of 77, max 552, total 10
1741, using 0/100000 read
iu@bielinux[mac_share] ls
hoge.fasta JSLAB_1_kadota.pdf uge
iu@bielinux[mac_share]
```

① ugeディレクトリに移動してls。② contigs.faが主なアセンブリ結果ファイルです

# 移動して確認

```
#解析したいファイル(hoge.fasta)のダウンロード  
wget -c http://www.iu.a.u-tokyo.ac.jp/%7Ekadota/20160912/hoge.fasta
```

```
#行数やファイルサイズを表示(行数が200,000行、サイズが6,688,895 bytesとなっていればOK)  
wc hoge.fasta
```

```
#最初の10行分を表示(2行分で1つのリードを表し、1リードが50塩基であることを確認)  
head hoge.fasta
```

```
#de novoアセンブリを実行し、結果をugeディレクトリに保存  
velveth uge 31 -short -fasta hoge.fasta  
velvetg uge
```

```
#ugeディレクトリに移動して、アセンブリ結果ファイル(contigs.fa)があることを確認
```

```
cd uge  
ls
```

```
#行数やファイルサイズを表示(入力のhoge.fasta)  
wc contigs.fa
```

```
#最初の10行分を表示(アセンブリ結果として、  
head contigs.fa
```

The terminal window shows the following output:

```
[1.271936] Writing contigs into uge/contigs.fa...  
[1.288760] Writing into stats file uge/stats.txt...  
[1.299793] Writing into graph file uge/LastGraph...  
Final graph has 1259 nodes and n50 of 77, max 552, total 10  
1741, using 0/100000 reads  
iu@bielinux[mac_share] ls [ 4:25午後 ]  
hoge.fasta JSLAB 1 kadota.pdf uge  
iu@bielinux[mac_share] cd uge [ 4:30午後 ]  
iu@bielinux[uge] ls [ 4:38午後 ]  
contigs.fa LastGraph PreGraph Sequences  
Graph Log Roadmaps stats.txt  
iu@bielinux[uge] [ 4:38午後 ]
```

Red arrows and boxes highlight the `ls` command in the `uge` directory, which shows `contigs.fa` as the primary assembly result file.

# WCで確認

①wcでアセンブリ結果ファイル(contigs.fa)の行数を確認。②4,038行。入力(hoge.fasta)は200,000行であることから、行数が大幅に減ったことがわかる

```
#解析したいファイル(hoge.fasta)のダウンロード
wget -c http://www.iu.a.u-tokyo.ac.jp/%7Ekadota/20160912/hoge.fasta

#行数やファイルサイズを表示(行数が200,000行、サイズが6,688,895 bytesとなっていればOK)
wc hoge.fasta

#最初の10行分を表示(2行分で1つのリードを)
head hoge.fasta

#de novoアセンブリを実行し、結果をugeディレクトリに保存
velveth uge 31 -short -fasta hoge.fasta
velvetg uge

#ugeディレクトリに移動して、アセンブリ結果を確認
cd uge
ls

#行数やファイルサイズを表示(入力のhoge.fastaと比較)
wc contigs.fa

#最初の10行分を表示(アセンブリ結果として)
head contigs.fa
```

```
iu@bielinux[uge] pwd [ 5:03午後 ]
/home/iu/Desktop/mac_share/uge
iu@bielinux[uge] ls [ 5:03午後 ]
contigs.fa LastGraph PreGraph Sequences
Graph Log Roadmaps stats.txt
iu@bielinux[uge] wc contigs.fa [ 5:03午後 ]
4038 4038 183069 contigs.fa [ 5:03午後 ]
iu@bielinux[uge] [ 5:03午後 ]
```

# headで確認

①headでアセンブリ結果ファイル(contigs.fa)の最初の10行分を表示。パッと見て、入力(50塩基の長さのリードが100,000個)よりも長い塩基配列(コンティグという)が得られていることがわかる

#解析したいファイル(hoge.fasta)のダウンロード  
wget -c http://www.iu.a.u-tokyo.ac.jp/%7Ekadota/20160912/hoge.f

#行数やファイルサイズを表示(行数が200,000行、サイズが6,688,895 bytesとなっていればOK)  
wc hoge.fasta

#最初の10行分を表示(2行分で1つのリードを)  
head hoge.fasta

#de novoアセンブリを実行し、結果をugeディレクトリに保存  
velveth uge 31 -short -fasta hoge.fasta  
velvetg uge

#ugeディレクトリに移動して、アセンブリ結果を確認  
cd uge  
ls

#行数やファイルサイズを表示(入力のhoge.fasta)  
wc contigs.fa

#最初の10行分を表示(アセンブリ結果として)  
head contigs.fa

```
iu@bielinux[uge] pwd [ 5:03午後 ]
/home/iu/Desktop/mac_share/uge
iu@bielinux[uge] ls [ 5:03午後 ]
contigs.fa LastGraph PreGraph Sequences
Graph Log Roadmaps stats.txt
iu@bielinux[uge] wc contigs.fa [ 5:03午後 ]
4038 4038 183069 contigs.fa
iu@bielinux[uge] head contigs.fa [ 5:03午後 ]
>NODE_1_length_124_cov_1.596774
CGGATTAGCTCAATTGATTATGGGAAATGCTGTACCCGGAGCCGTA
```

# Contents

## ■ イン트로ダクション

- 概要、背景(NGS用カリキュラム、講習会)、Linuxスキル習得の意義
- ウェブ情報(日本乳酸菌学会誌のNGS連載やNGS講習会資料)

## ■ 実習環境に慣れる

- 仮想環境での作業に慣れる
- GUIとCUI(マウス操作かコマンド入力操作か)
- ターミナルでの作業
- 共有フォルダの概念を理解

## ■ 練習

- 作業ディレクトリの変更、練習用NGSデータファイルのダウンロード
- ファイルの確認、de novoゲノムアセンブリ
- BLAST検索

## ■ 課題

- グループごとに異なる課題ファイルを入力として、「ダウンロード、de novoアセンブリ、BLAST検索」を実行し、得られた結果をレポートにまとめて発表せよ
- グループ1はkadai1.fasta、グループ2はkadai2.fasta、etc.

# BLAST検索

世界中から得られた塩基配列のデータベース(の一部)に対して、手元にある塩基配列をBLASTというプログラムを用いて検索する作業。配列相同性検索ともいいます。詳細については秋の講義科目「生物情報科学」で説明がなされると思います。ここでは詳細はすっ飛ばして、必要最小限の作業を行う

・ 門田幸二「講義資料」, 東京大学農学部展開科目: バイオイン

内容: バイオインフォマティクスやNGSをキーワード程度は知っ  
オインフォマティクスとLinuxスキル習得の意義。バイオインフ  
るアンケート結果)。バイオインフォマティクス人材育成のための講習会  
めに必要な基本スキルの習得。具体的には、日本乳酸菌学会誌のNGS連載内容を理解するための基礎として、[Bio-Linux](#)のノリに慣  
れるのが最低限の到達目標。何かやったという達成感を味わってもらうため、NGSデータのde novoゲノムアセンブリを行い、元のリー  
ド長(塩基配列の長さ)よりも伸びたことを実感してもらう。また、アセンブリ結果の一部をBLAST検索し、入力データ(hoge.fasta; 約  
6.4MB)がどんな生物種であったかを調べるあたりまで。課題はkadai1.fasta, kadai2.fasta, kadai3.fasta, kadai4.fastaのいずれかを選択  
して実行。下記と同様の手順で「データのダウンロード → de novoアセンブリ → BLAST検索」の作業を行い、得られた結果を発表。約2日分。

#作業ディレクトリの変更

```
cd /home/iu/Desktop/mac_share
```

#解析したいファイル(hoge.fasta)のダウンロード

```
wget -c http://www.iu.a.u-tokyo.ac.jp/%7Ekadota/20160912/hoge.fasta
```

#行数やファイルサイズを表示(行数が200,000行、サイズが6,688,895 bytesとなっていればOK)

```
wc hoge.fasta
```

#最初の10行分を表示(2行分で1つのリードを表し、1リードが50塩基であることを確認)

```
head hoge.fasta
```

#de novoアセンブリを実行し、結果をugeディレクトリに保存

```
velveth uge 31 -short -fasta hoge.fasta
```

```
velvetg uge
```

#ugeディレクトリに移動して、アセンブリ結果ファイル(contigs.fa)があることを確認

```
cd uge
```

```
ls
```

#行数やファイルサイズを表示(入力のhoge.fastaは200,000行だったが、出力のcontigs.faは4,038行となっている)

```
wc contigs.fa
```

BLASTのトップ画面。①の部分にアセンブリ結果として得られた配列の一部を入力としてBLASTを実行する

# BLAST検索

http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE\_TYPE=Blas Nucleotide BLAST: Sear... x

NIH U.S. National Library of Medicine NCBI Sign in to NCBI

BLAST® >> blastn suite Home Recent Results Saved Strategies Help

Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#) [Reset page](#) [Bookmark](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) Query subrange [From](#)  [To](#)

Or, upload file  [Job Title](#)  [Enter a descriptive title for your BLAST search](#)

Align two or more sequences

Choose Search Set

Database  Human genomic + transcript  Mouse genomic + transcript  Others (nr etc.):  
Nucleotide collection (nr/nt)  [Organism](#)   Exclude

# BLAST検索

アセンブリ実行結果の、①最初のコンティグ(反転部分)をコピー

The image shows a web browser window displaying the NCBI BLAST search interface. The browser address bar shows the URL: `http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=Blas`. The page title is "Nucleotide BLAST: Search". The interface includes a search form with fields for "Enter Query Sequence", "Enter accession number(s), gi(s), or FASTA", "Job Title", and "Choose Search Set". The "Choose Search Set" section has "Database" set to "Human genomic + Nucleotide collection" and "Organism" set to "Optional".

Overlaid on the browser is a terminal window titled `iu@bielinux[~/Desktop/mac_share/uge]`. The terminal shows the following commands and output:

```
iu@bielinux[uge] pwd [ 5:03午後 ]
/home/iu/Desktop/mac_share/uge
iu@bielinux[uge] ls [ 5:03午後 ]
contigs.fa LastGraph PreGraph Sequences
Graph Log Roadmaps stats.txt
iu@bielinux[uge] wc contigs.fa [ 5:03午後 ]
4038 4038 183069 contigs.fa
iu@bielinux[uge] head contigs.fa [ 5:03午後 ]
>NODE_1_length_124_cov_1.596774
CGGATTAGCTCAATTGATTATGGGAAATGCTGTACCC
GGGTAAGGTGTTGATGAACTTGGCTGGACCAAACA
AATCTTGCTATTTGTCTTGAGTGGGTTCTTCCGC
>NODE_2_length_81_cov_1.530864
ATCCAACCGCACCAAAGTTGGCTGTCATGACTATCCG
CTTCAAGATGCGAATAATACCTTTTTTGCAGCCTG
>NODE_3_length_94_cov_2.340425
TGGAGTTGTAATAATGGCAAATAATAATGATTCTACT
CAGTAAAAAGAATCAACAACAAAATGAACAAAACCAATCCGGACCAGTTAAATCAAAGAA
iu@bielinux[uge] [ 5:16午後 ]
```

A context menu is open over the terminal output, with the following options: "Open Terminal", "Open Tab", "Close Window", "Copy", "Paste", "Profiles", and "Show Menubar". The "Copy" option is highlighted with a red circle containing the number 1.

# BLAST検索

①赤枠内でペースト。これが②問い合わせしたい塩基配列 (Query Sequence) です。③ページ下部にスクロール

The screenshot shows the NCBI BLAST search page. The browser address bar shows the URL: `http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=Blas`. The page title is "Nucleotide BLAST: Search...". The navigation bar includes "NIH", "U.S. National Library of Medicine", "NCBI", and "Sign in to NCBI". The main heading is "BLAST >> blastn suite" with links for "Home", "Recent Results", "Saved Strategies", and "Help". The sub-heading is "Standard Nucleotide BLAST".

Annotations on the page:

- ①: A red arrow points to a red-bordered box containing a FASTA sequence: `>NODE_1_length_124_cov_1.596774  
CGGATTAGCTCAATTGATTATGGGAAATGCTGTACCCGGAGCCGTA  
CTTGGTATTTTAGT  
GGGTAAGGTGTTGATGA  
ACTTGGCTGGACCAAA  
CAACTAAAATTATGTT  
AACAGCTGT  
AATCTTGCTATTTGT  
CTTGAGTGGGTTCTT  
CCGC`
- ②: A red arrow points to the "tblastn" tab in the "BLASTN programs search nucleotide databases using a nucleotide query. more..." section.
- ③: A red arrow points to the right side of the page, indicating the need to scroll down to view results.

Other visible elements include the "Enter Query Sequence" section with a "Clear" button, a "Query subrange" section with "From" and "To" input fields, and a "Choose Search Set" section with radio buttons for "Human genomic + transcript", "Mouse genomic + transcript", and "Others (nr etc.)", and a dropdown menu for "Nucleotide collection (nr/nt)".

# BLAST検索

http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE\_TYPE=Blas Nucleotide BLAST: Sear...

Nucleotide collection (nr/nt)

Organism  
Optional

Enter organism name or id—completions will be suggested  Exclude +  
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude  
Optional

Models (XM/XP)  Uncultured/environmental sample sequences

Limit to  
Optional

Sequences from type material

Entrez Query  
Optional

[YouTube](#) [Create custom database](#)  
Enter an Entrez query to limit search

Program Selection

Optimize for

Highly similar sequences (megablast)  
 More dissimilar sequences (discontiguous megablast)  
 Somewhat similar sequences (blastn)  
Choose a BLAST algorithm

**BLAST** Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)  
 Show results in a new window

+ Algorithm parameters

BLAST is a registered trademark of the National Library of Medicine.

[Copyright](#) | [Disclaimer](#) | [Privacy](#) | [Accessibility](#) | [Contact](#) | [Send feedback](#) [NCBI](#) | [NLM](#) | [NIH](#) | [DHHS](#)

# BLAST検索

The screenshot shows a web browser window with the URL <http://blast.ncbi.nlm.nih.gov/Blast.cgi>. The page header includes the NIH logo, 'U.S. National Library of Medicine', and 'NCBI'. The main content area is titled 'BLAST' and shows the path 'blastn suite >> RID-VUSB37TU01R'. The specific page is 'Format Request Status' for job 'NODE\_1\_length\_124\_cov\_1.596774'. A table displays the following information:

Request ID	VUSB37TU01R
Status	Searching
Submitted at	Wed Aug 24 04:57:07 2016
Current time	Wed Aug 24 04:57:10 2016
Time since submission	00:00:02

Below the table, it states: 'This page will be automatically updated in 2 seconds'. The footer contains copyright information and links to 'NCBI | NLM | NIH | DHHS'.

# BLAST検索

サーバの混み具合にも依存しますが、概ね1分以内にこのような①BLAST実行結果が得られます。②問い合わせ配列は塩基配列で、長さは154塩基だったことがわかります。③検索対象のDB中にヒットした(一致した)ものが1つだけあったと解釈する。④ちょっと下のほうに移動

U.S. National Library of Medicine | NCBI | Sign in to NCBI

**BLAST** » **blastn suite** » RID-VUSB37TU01R

Home Recent Results Saved Strategies Help

BLAST Results

Edit and Resubmit Save Search Strategies Formatting options Download YouTube How to read this page Blast report des

**NODE\_1\_length\_124\_cov\_1.596774**

RID [VUSB37TU01R](#) (Expires on 08-25 16:57 pm)

Query ID	Id Query_100105	Database Name	nr
Description	NODE_1_length_124_cov_1.596774	Description	Nucleotide collection (nt)
Molecule type	nucleic acid	Program	BLASTN 2.5.0+ <a href="#">Citation</a>
Query Length	154		

Other reports: [Search Summary](#) [Taxonomy reports] [Distance tree of results]

**Graphic Summary**

Distribution of 1 Blast Hits on the Query Sequence

Mouse over to see the define, click to show alignments

Color key for alignment scores

Query	<40	40-50	50-80	80-200	>=200
-------	-----	-------	-------	--------	-------

# BLAST検索

①このあたり。検索対象のDB中でヒットしたものは、②*Lactobacillus hokkaidonensis* (ある乳酸菌株)の完全なゲノム配列

The screenshot shows the NCBI BLAST search results page. The browser address bar displays <http://blast.ncbi.nlm.nih.gov/Blast.cgi>. The page is titled "Graphic Summary" and "Descriptions".

**Graphic Summary:** A horizontal bar chart titled "Distribution of 1 Blast Hits on the Query Sequence". The x-axis represents the query sequence length from 1 to 150. A color key for alignment scores is provided: <40 (black), 40-50 (blue), 50-80 (green), 80-200 (magenta), and >=200 (red). The query sequence is shown as a red bar, indicating a score of >=200.

**Descriptions:** A table titled "Sequences producing significant alignments:" shows one result. The table has columns for Description, Max score, Total score, Query cover, E value, Ident, and Accession.

Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/> <a href="#">Lactobacillus hokkaidonensis DNA, complete genome</a>	285	285	100%	2e-73	100%	<a href="#">AP014680.1</a>



# BLAST検索

①さらにページ下部に移動。②Alignmentsというところ。154塩基の問い合わせ配列(Query)が、③乳酸菌ゲノム配列のどのあたりにヒットしたのかを並べて(alignmentして)示した結果。④乳酸菌ゲノム配列の全長は、2,277,985塩基(約2.3Mb;メガbaseの意味)

http://blast.ncbi.nlm.nih.gov/Blast.cgi

Alignments

Download ▾ GenBank Graphics

Lactobacillus hokkaidonensis DNA, complete genome  
Sequence ID: [dbj|AP014680.1](#) Length: 2277985 Number of Matches: 1

Range 1: 583259 to 583412 [GenBank](#) [ORA](#)

Score	Expect	Identities	Gaps	Strand
285 bits(154)	2e-73	154/154(100%)	0/154(0%)	Plus/Plus
Query 1	CGGATTAGCTCAATTGATTATGGGAAATGCTGTACCCGGAGCCGTA	CGGATTAGCTCAATTGATTATGGGAAATGCTGTACCCGGAGCCGTA		60
Sbjct 583259	CGGATTAGCTCAATTGATTATGGGAAATGCTGTACCCGGAGCCGTA	CGGATTAGCTCAATTGATTATGGGAAATGCTGTACCCGGAGCCGTA		583318
Query 61	GGGTAAAGGTGTTGATGAACTTGGCTGGACAAAACAATAAATTATGTTAACAGCTGT	GGGTAAAGGTGTTGATGAACTTGGCTGGACAAAACAATAAATTATGTTAACAGCTGT		120
Sbjct 583319	GGGTAAAGGTGTTGATGAACTTGGCTGGACAAAACAATAAATTATGTTAACAGCTGT	GGGTAAAGGTGTTGATGAACTTGGCTGGACAAAACAATAAATTATGTTAACAGCTGT		583378
Query 121	AATCTTGCTATTTGCTTGAGTGGGTTCTTCCGC	AATCTTGCTATTTGCTTGAGTGGGTTCTTCCGC		154
Sbjct 583379	AATCTTGCTATTTGCTTGAGTGGGTTCTTCCGC	AATCTTGCTATTTGCTTGAGTGGGTTCTTCCGC		583412

Related Information

# BLAST検索

①154塩基の問い合わせ配列(Query sequence)が上、②ヒットした乳酸菌ゲノム配列 (Subject sequenceの略でSbjct)が下

Alignments

Download ▾ GenBank Graphics ▾ Next ▲ Previous ▲ Descriptions

Lactobacillus hokkaidonensis DNA, complete genome  
Sequence ID: [dbj|AP014680.1](#) Length: 2277985 Number of Matches: 1

Range 1: 583259 to 583412 GenBank Graphics ▾ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand	
285 bits(154)	2e-73	154/154(100%)	0/154(0%)	Plus/Plus	
Query 1	CGGATTAGCTCAATTGATTATGGGAAATGCTGTACCCGGAGCCGTA	CGGATTAGCTCAATTGATTATGGGAAATGCTGTACCCGGAGCCGTA	CGGATTAGCTCAATTGATTATGGGAAATGCTGTACCCGGAGCCGTA	CGGATTAGCTCAATTGATTATGGGAAATGCTGTACCCGGAGCCGTA	60
Sbjct 583259	CGGATTAGCTCAATTGATTATGGGAAATGCTGTACCCGGAGCCGTA	CGGATTAGCTCAATTGATTATGGGAAATGCTGTACCCGGAGCCGTA	CGGATTAGCTCAATTGATTATGGGAAATGCTGTACCCGGAGCCGTA	CGGATTAGCTCAATTGATTATGGGAAATGCTGTACCCGGAGCCGTA	583318
Query 61	GGGTAAAGGTGTTGATGAACTTGGCTGGACAAAACAATAAAATTATGTTAACAGCTGT	GGGTAAAGGTGTTGATGAACTTGGCTGGACAAAACAATAAAATTATGTTAACAGCTGT	GGGTAAAGGTGTTGATGAACTTGGCTGGACAAAACAATAAAATTATGTTAACAGCTGT	GGGTAAAGGTGTTGATGAACTTGGCTGGACAAAACAATAAAATTATGTTAACAGCTGT	120
Sbjct 583319	GGGTAAAGGTGTTGATGAACTTGGCTGGACAAAACAATAAAATTATGTTAACAGCTGT	GGGTAAAGGTGTTGATGAACTTGGCTGGACAAAACAATAAAATTATGTTAACAGCTGT	GGGTAAAGGTGTTGATGAACTTGGCTGGACAAAACAATAAAATTATGTTAACAGCTGT	GGGTAAAGGTGTTGATGAACTTGGCTGGACAAAACAATAAAATTATGTTAACAGCTGT	583378
Query 121	AATCTTGCTATTTGCTTGAGTGGGTTCTTCCGC	AATCTTGCTATTTGCTTGAGTGGGTTCTTCCGC	AATCTTGCTATTTGCTTGAGTGGGTTCTTCCGC	AATCTTGCTATTTGCTTGAGTGGGTTCTTCCGC	154
Sbjct 583379	AATCTTGCTATTTGCTTGAGTGGGTTCTTCCGC	AATCTTGCTATTTGCTTGAGTGGGTTCTTCCGC	AATCTTGCTATTTGCTTGAGTGGGTTCTTCCGC	AATCTTGCTATTTGCTTGAGTGGGTTCTTCCGC	583412

# BLAST検索

154塩基の問い合わせ配列(Query sequence)の①1塩基目から②154塩基目が、②ヒットした乳酸菌ゲノム配列(Subject sequenceの略でSbjct)の③583,259塩基目から④583,412塩基目の領域で、⑤完全一致していたことがわかる

Alignments

Download ▾ GenBank Graphics

Lactobacillus hokkaidonensis DNA, complete genome  
Sequence ID: [dbj|AP014680.1](#) Length: 2277985 Number of Matches: 1

Range 1: 583259 to 583412 GenBank Graphics

Score	Expect	Identities	Gaps	Strand
285 bits	2e-73	154/154(100%)	0/154(0%)	Plus/Plus
Query 1	CGGATTAGCTC			60
Sbjct 583259	CGGATTAGCTCAATTGATTATGGGAAATGCTGTACCCGGAGCCGTA			583318
Query 1	GGGTAAAGGTGTTGATGAACTTGGCTGGACCAAAACA			120
Sbjct 583319	GGGTAAAGGTGTTGATGAACTTGGCTGGACCAAAACA			583378
Query 121	AATCTTGCTATTTGCTTGAGTGGGTTCTTCCGC			154
Sbjct 583379	AATCTTGCTATTTGCTTGAGTGGGTTCTTCCGC			583412

# Contents

## ■ イントロダクション

- 概要、背景(NGS用カリキュラム、講習会)、Linuxスキル習得の意義
- ウェブ情報(日本乳酸菌学会誌のNGS連載やNGS講習会資料)

## ■ 実習環境に慣れる

- 仮想環境での作業に慣れる
- GUIとCUI(マウス操作かコマンド入力操作か)
- ターミナルでの作業
- 共有フォルダの概念を理解

## ■ 練習

- 作業ディレクトリの変更、練習用NGSデータファイルのダウンロード
- ファイルの確認、de novoゲノムアセンブリ
- BLAST検索

## ■ 課題

- グループごとに異なる課題ファイルを入力として、「ダウンロード、de novoアセンブリ、BLAST検索」を実行し、得られた結果をレポートにまとめて発表せよ
- グループ1はkaday1.fasta、グループ2はkaday2.fasta、etc.

# 課題

・ 門田幸二「[講義資料](#)」, 東京大学農学部展開科目: バイオインフォマティクス, 東京大学(東京), 2016.09.12-16

内容: バイオインフォマティクスやNGSをキーワード程度は知っているヒト(学部3年生程度)向けのアクティブ・ラーニング系講義。バイオインフォマティクスとLinuxスキル習得の意義。バイオインフォマティクス分野で需要が多いのはNGSデータの解析 (JST-NBDCIによる[アンケート結果](#))。バイオインフォマティクス人材育成のための講習会の一環として行われているNGS講習会資料を有効活用するために必要な基本スキルの習得。具体的には、日本乳酸菌学会誌のNGS連載内容を理解するための基礎として、[Bio-Linux](#)のノリに慣れるのが最低限の到達目標。何かやったという達成感を味わってもらうため、NGSデータのde novoゲノムアセンブリを行い、元のリード長(塩基配列の長さ)よりも伸びたことを実感してもらう。また、アセンブリ結果の一部をBLAST検索し、入力データ([hoge.fasta](#); 約6.4MB)がどんな生物種であったかを調べるあたりまで。課題は[kadai1.fasta](#), [kadai2.fasta](#), [kadai3.fasta](#), [kadai4.fasta](#)のいずれかを選択して実行。下記と同様の手順で「データのダウンロード → de novoアセンブリ → BLAST検索」し、得られた結果表。約2日分。

#作業ディレクトリの変更

```
cd /home/iu/Desktop/mac_share
```

#解析したいファイル([hoge.fasta](#))のダウンロード

```
wget -c http://www.iu.a.u-tokyo.ac.jp/%7Ekladota/20160912/hoge.fasta
```

#行数やファイルサイズを表示(行数が200,000行、サイズが6,688,895 bytesとなっていればOK)

```
wc hoge.fasta
```

#最初の10行分を表示(2行分で1つのリードを表し、1リードが50塩基であることを確認)

```
head hoge.fasta
```

#de novoアセンブリを実行し、結果をugeディレクトリに保存

```
velveth uge 31 -short -fasta hoge.fasta
```

```
velvetg uge
```

#ugeディレクトリに移動して、アセンブリ結果ファイル(contigs.fa)があることを確認

```
cd uge
```

```
ls
```

#行数やファイルサイズを表示(入力の[hoge.fasta](#)は200,000行だったが、出力のcontigs.faは4,038行となっている)

```
wc contigs.fa
```



# 実習用PC環境を自力で

## (Rで)塩基配列解析

(last modified 2016/08/12, since 2011)

このウェブページは [インストール](#) についての推奨手順 (Windows2015.04.04版とMacintosh2015.0

に従ってフリーソフトRと必要なパッケージをインストール済であるという前提を記述しています。

の方は [基本的な利用法](#) (Win  
ページを体系的にまとめた書

### What's new?

- [NGSハンズオン講習会2016](#) (通すところのミス、スラッシュ実行もうまくいく)あたりです。
- [NGSハンズオン講習会2016](#)
- [NGSハンズオン講習会2016](#)
- [Twitterハッシュタグ AJAC](#)
- [NGSハンズオン講習会2016](#) データロード可能性は本記事

- 書籍 | [トランスクリプトーム解析 | 4.3.2 データの正規化\(応用編\)](#) (last modified 2014/04/27)
- 書籍 | [トランスクリプトーム解析 | 4.3.3 2群間比較](#) (last modified 2014/04/28)
- 書籍 | [トランスクリプトーム解析 | 4.3.4 他の実験デザイン\(3群間\)](#) (last modified 2014/04/28)
- 書籍 | [日本乳酸菌学会誌 | 第6回ゲノムアセンブリ](#) (last modified 2016/07/01)
- 書籍 | [日本乳酸菌学会誌 | 第1回イントロダクション](#) (last modified 2015/09/11)
- 書籍 | [日本乳酸菌学会誌 | 第2回GUI環境からコマンドライン環境へ](#) (last modified 2015/11/26)
- 書籍 | [日本乳酸菌学会誌 | 第3回Linux環境構築からNGSデータ取得まで](#) (last modified 2015/12/07)
- 書籍 | [日本乳酸菌学会誌 | 第4回クオリティコントロールとプログラムのインストール](#) (last modified 2016/07/05)
- 書籍 | [日本乳酸菌学会誌 | 第5回アセンブル、マッピング、そしてQC](#) (last modified 2016/07/05)
- 書籍 | [日本乳酸菌学会誌 | 第6回ゲノムアセンブリ](#) (last modified 2016/06/17)
- 書籍 | [日本乳酸菌学会誌 | 第7回ロングリードアセンブリ](#) (last modified 2016/07/01)

### 書籍 | 日本乳酸菌学会誌 | 第6回ゲノムアセンブリ

日本乳酸菌学会誌の第6回分です。Linuxコマンドのリンク先は主に [日経BP社](#) 様です。原稿PDFの「はじめに」項目のところにtypoがあります。誤:するであれば、正:する"の"であれば、ですねm( )m。また、「配列長によるフィルタリング」項目の最後の文章「これらについては、第7回で詳述する予定である。」についてですが、これは「これらについては、"第8回以降"で詳述する予定である。」と読み替えてください。第7回ドラフト原稿作成時点で、これらの内容は含まれていないためです(2016年4月23日追加)。

- [原稿PDF](#)
- ウェブ資料PDF
  - [Windows用](#) (2016.06.17版; 約20MB)
  - [Macintosh用](#)
- (共有フォルダ設定情報を含む)連載第3回終了時点以降の ova ファイルからのインストール手順:
  - [Windows用](#) (2015.12.28版; 約2MB)
  - [Macintosh用](#) (2015.12.28版; 約2MB)

実習用PCは、既にVirtualBoxをインストールし、Bio-Linuxというものを導入(インポート)し、共有フォルダを設定した後の状態です。この環境を自力で構築したいヒトは、①第6回ゲノムアセンブリ、②のインストール手順を参考にしてください。

# 実習用PC環境を自力で

書籍 | 日本乳酸菌学会誌 | 第6回ゲノムアセンブリ

実習用PCと完全に同じ環境にしたいヒトは、①のスライド5のところで私宛にメールする際に、件名を「乳酸菌連載第4回終了時点のovaファイル希望」としてください

日本乳酸菌学会誌の第6回分です。Linuxコマンドのリンク先は主に日経BP社様です。原稿PDFの「はじめに」項目のところにtypoがあります。誤:するであれば、正:するの"であれば、です。ねm( )m。また、「配列長によるフィルタリング」項目の最後の文章「これらについては、第7回で詳述する予定である。」についてですが、これは「これらについては、"第8回以降"で詳述する予定である。」と読み替えてください。第7回ドラフト原稿作成時点で、これらの内容は含まれていないためです(2016年4月23日追加)。

- [原稿PDF](#)
- ウェブ資料PDF
  - [Windows用](#)(2016.06.17版; 約20MB)
  - [Macintosh用](#)
- (共有フォルダ設定情報を含む)連載第3回終了時点以降のovaファイルからのインストール手順:

- [Windows用](#)(2015.12.28版; 約2MB)
- [Macintosh用](#)(2015.12.28版; 約2MB)

## ovaファイルの準備

### (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 2015/12/26, since 2011)

まずは①赤下線の指示通りに本文は空で②メール送信。しばらくすると(門田がオンラインなら概ね数時間程度以内に)Google driveのURL情報が返信される。

What's new?

- このウェブページはインストール | についての従ってフリーソフト R と必要なパッケージをインストール | 基本的な利用法 (Windows 2015.04.03 版と Mac 的にまとめた書籍 | もあります。(2015/04/03)
- 多群間比較用の推奨ガイドライン提唱論文 | 概要については | 門田 | のページでも紹介して | ら DEG 検出結果のおおよその見種も | が可能 | 推奨ガイドライン周辺の関連項目もアップデート
- [日本乳酸菌学会誌の NGS 関連連載の第 5 回 | したので、若干プログラムのバージョンが | 上か](#)
- [解析 | 一般 | アライメント | についてを追加](#)
- [日本乳酸菌学会誌の NGS 関連連載の第 4 回 | フェブ資料と更新 | しました。2015 年 12 月初旬に | 一新 | したので、若干プログラムのバージョンが | 上が | っています。各回終了時点の ova ファイル \(約 6 GB\) も提供 | 可能 | です。\(権利関係上無条件公開は | できませんので...\) | 欲しい方は、メールの | タイトルを | 「乳酸菌連載第 x 回 | 終了 | 時点の ova ファイル希望」として | 私宛に | メール | してください \(本文は空で OK\)。URL を | お知らせ | します。\(2015/12/11\) NEW](#)

日本乳酸菌学会誌の連載第6回

5