生命科学領域における情報科学、統計科学の人材育成実践とその考察

石井一夫（東京農工大学）

生命科学領域の統計教育の特徴

- 近年、情報処理システム、大規模なデータ産生機器の普及などの伴い統計に対する需要が増えている。

- もともと、化学/生物などを得意とする学生、教員が多く、数理統計教育、情報教育は強くない傾向にある。

東京農工大学での取り組み

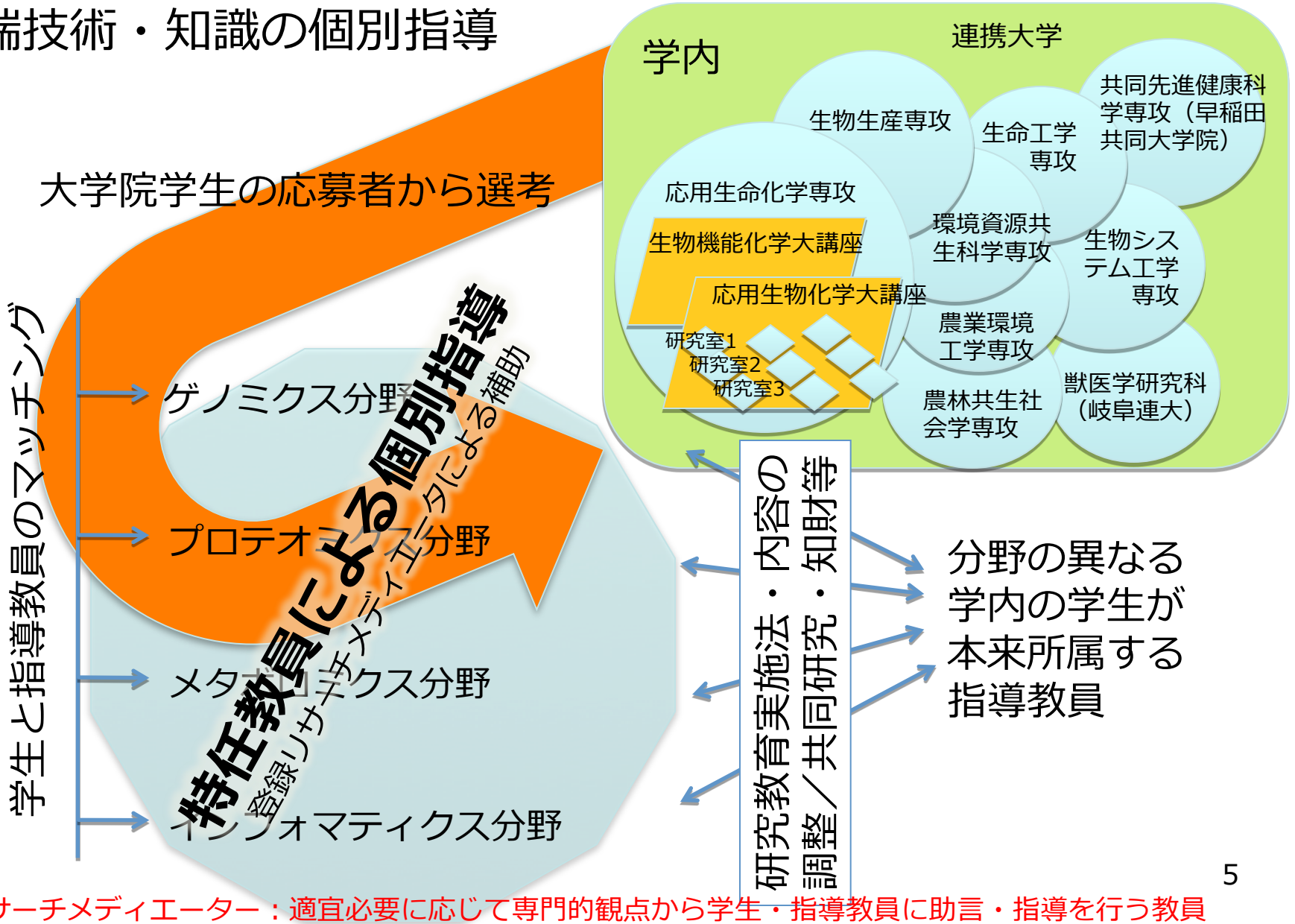
- 最近、統計科学、情報科学教育の需要の多いゲノム科学分野での人材育成プログラムにおいて
- 統計科学教育、情報科学教育に関する個別指導、セミナーなどを実施。
- ゲノム科学人材育成プログラム（H23～H27）文部科学省特別経費による

ゲノム科学人材育成プログラム

- 大学院生から卒業研究の研究課題を公募し、その中からゲノム科学、データ分析を希望する学生に対し、

- データ分析、統計科学、プログラミングなどの実践指導。

専攻・講座・研究教育分野／研究室の枠を超えた先端技術・知識の個別指導



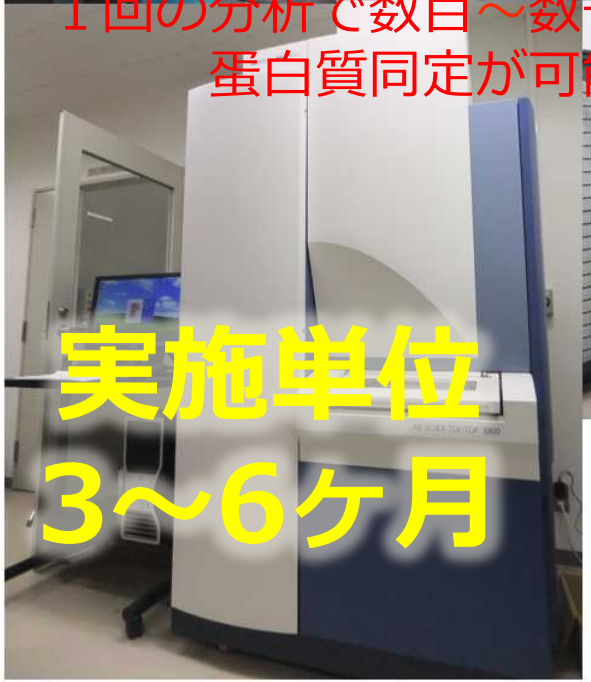
リサーチメディエーター：適宜必要に応じて専門的観点から学生・指導教員に助言・指導を行う教員

参考資料

最新のゲノム科学技術及びその応用技術と知識を習得

高性能質量分析装置
LTQ Orbitrap XL
ABSciex5800
Mascot解析システム
１回の分析で数百～数千種類の
蛋白質同定が可能

次世代ゲノムアナライザー
イルミナGAIIx/MiSeq/半導体
シークエンサー
１回の分析でDNA塩基3億～
300億個の配列決定が可能



実施単位
3~6ケ月

Linuxベースの大規模データ
解析システム等（６セット）

講習会用パソコンMacBook Air

ゲノム科学人材育成プログラム

- 3ヶ月ごとに1ステップで、一つの項目を実施。

- データ分析の指導

例：データの前処理、正規化、数値データの抽出、統計解析、データマイニング、機械学習など。

- 情報科学の指導

例：プログラミング（コーディング）、アルゴリズム

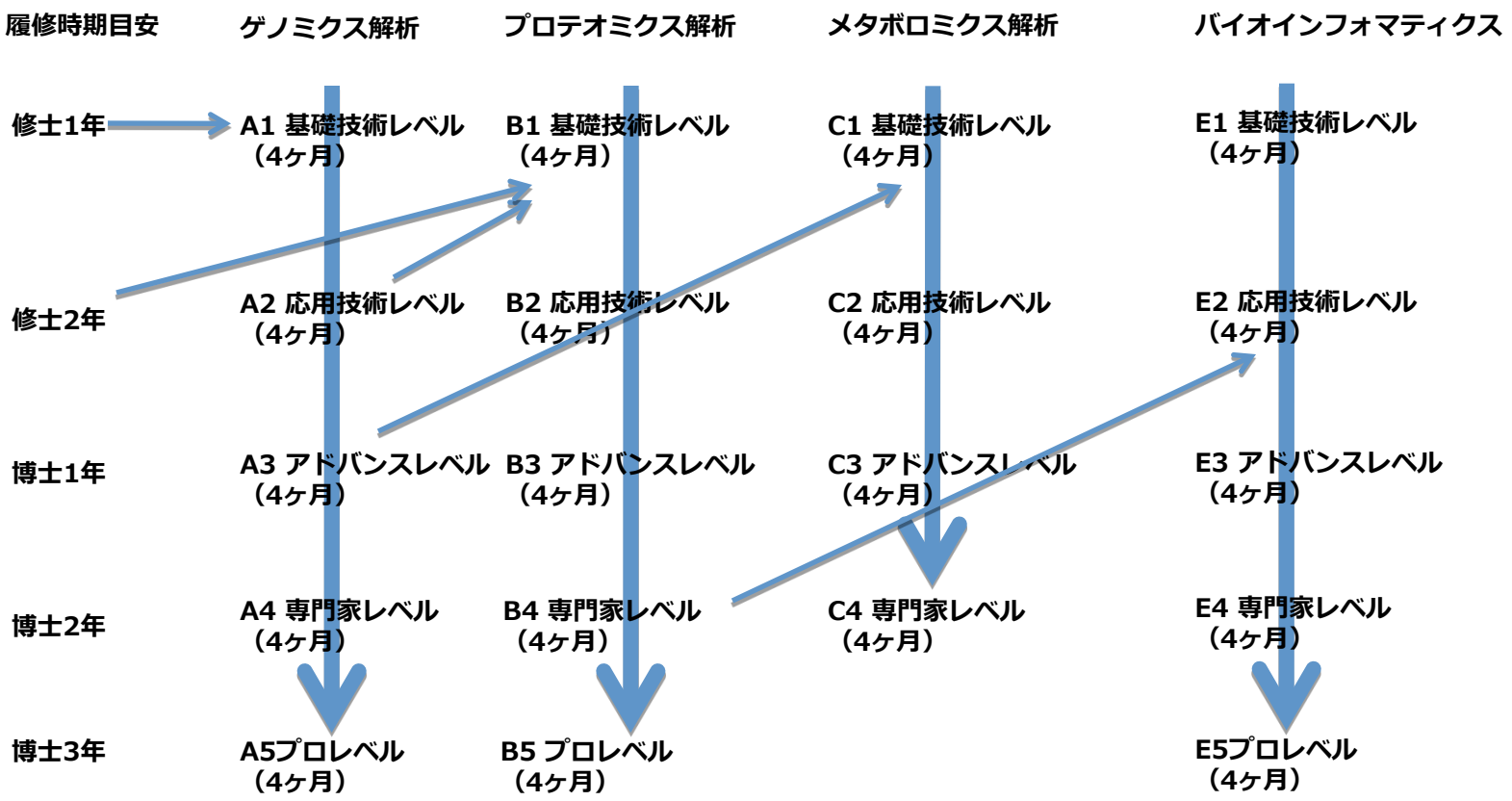
東京農工大学

提供する支援レベル（習得技術・内容）

基礎技術レベル（3ヶ月）	**E1:UNIXの操作・データ解析環境の立ち上げ・スクリプト作成（Perl/Ruby/Python）** **Linux の操作、インストール、Perlなどを用いたテキスト操作**
応用技術レベル（3ヶ月）	**E2:DNA配列アセンブリ・メタゲノム解析・データベース構築（SQL）** **各種アプリケーションソフトウェアの操作、データベースの操作と構築、簡単な統計解析**
アドバンスレベル（3ヶ月）	**E3:RNA-Seq解析・ChIP-Seq解析・統計解析（R/MatLab）** **統計解析、多変量解析、機械学習、クラスタ解析、グラフィックスによる視覚化**
専門家レベル（3ヶ月）	**E4:上記以外のデータ解析法（QTL・カスタムライブラリの解析）** **統計モデリング、モンテカルロシミュレーション**
プロレベル（3ヶ月）	**E5:新規データ解析法の開発実装（C/C++/Java）** **新規アルゴリズムの構築と実装**

東京農工大学

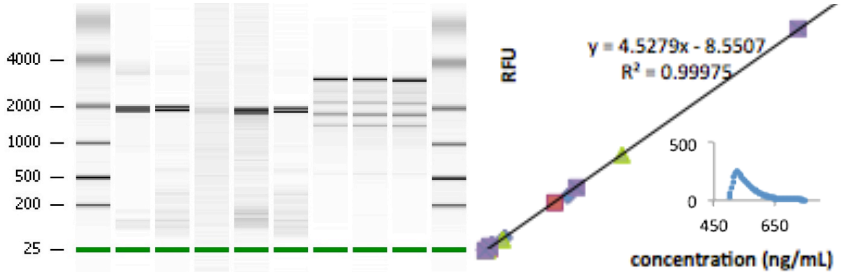
履修の進め方（例）



履修方法としては、解析技術毎に履修のレベルを段階的に上げて行くことを基本とする
例えば、修士2年にゲノミクス解析基礎技術レベルを習得した場合、博士1年に進んでゲノミクス解析応用技術レベル、アドバンスレベル、専門家レベルへとステップアップする。ゲノミクス解析アドバンスレベルからプロテオミクス解析に進みたい時には原則として基礎技術レベルから始める。どの学年から始める場合でも、原則基礎技術レベルから開始する。但し、既に経験が有る場合には、応用技術レベル、アドバンスレベルから開始することも可能とする。

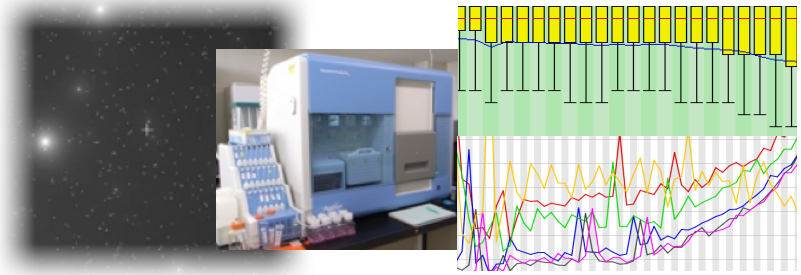
ゲノミクス解析・インフォマティクス 実施例

A1･E1: 基礎技術レベル



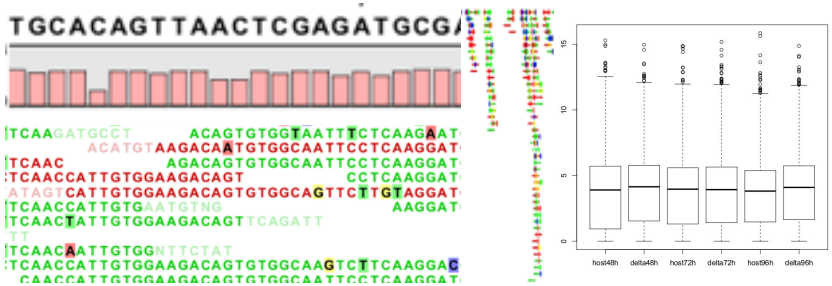
例 ゲノム試料の評価

A2･E2: 応用技術レベル



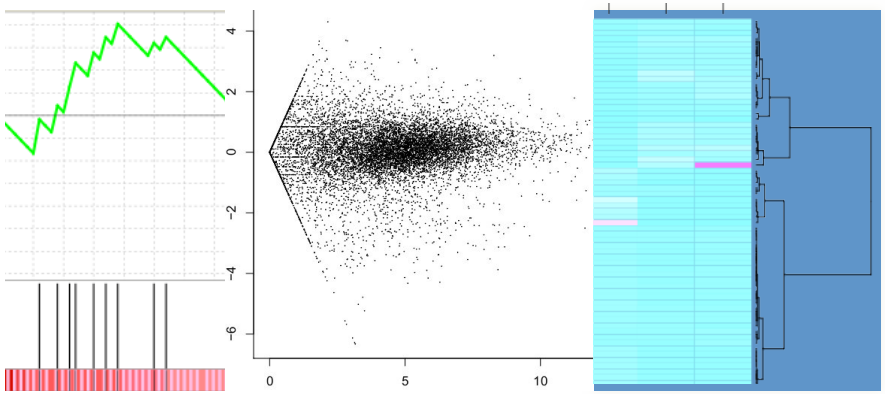
例 ゲノムアナライザーの稼働

B3: アドバンスレベル



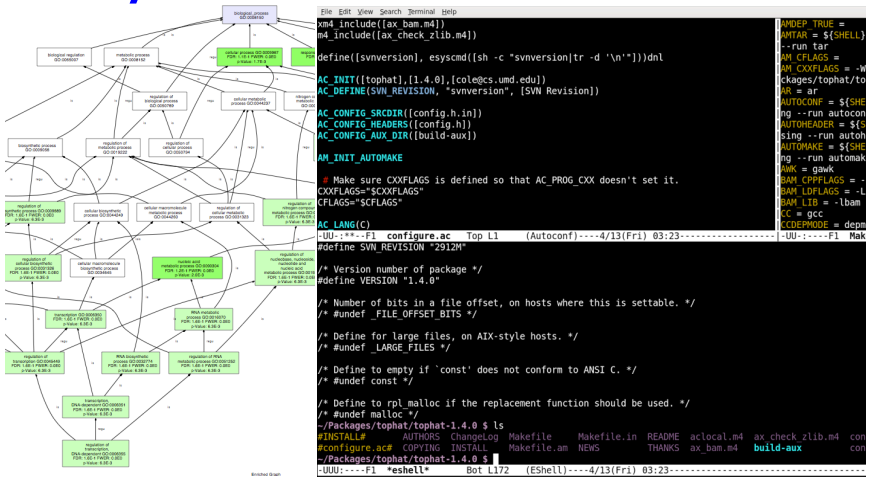
例 小型ゲノム解析

A4･E3: 専門家レベル



例 パスウェイ・トランスクリプトーム解析

A5･E4/E5: プロレベル



例 オントロジー解析、新規解析法の実装

・ゲノム科学関連企業32社、ゲノム解析で世界最大機関の北京ゲノム研究所などの海外研究機関との連携実績

2012年(平成24年)12月18日 火曜日 日刊工業新

課題に挑む 技術士のソリューション 〔139〕

東京農工大学大学院農学系ゲノム科学人材育成プログラム特任教授

石井 一夫（生物工学部門）

技術者教育㉒ 国際展開するゲノム人材育成

中国・BGIとセミナー 最先端研究を公開

世界最大規模

ゲノム科学分野では研究成果の発表や研究者間の情報交換も国際的に展開されるのが通常で、グローバル人材の育成を視野に入れた人材育成は必然だと言える。

BGIは、中国の深圳にあり03年に設立された世界最大規模のゲノム解析センターで、データ解析担当者も1500人以上いる。現在、ヒトゲノムをはじめ、イネ、カイコ、パンダ、ニワトリなどの世界をリードする動植物のゲノム解析を行ってきた。このたびゲノム科学人材育成の一環として、

カキゲノム解析

9月26日にBGIの研究現場で活躍している現役研究者を招聘し、大学内の大学院生向けにセミナーを企画した。同時に、学外の一般参加からの聴講者も受け入れた。

阿セミナーは、10月に英ネイチャー誌に掲載されたカキのゲノム解析にはじまり、遺伝子発現解析を利用した進化論的な系統樹による分析、ヒトのがん患者検体における遺伝子発現分析、トランスクリプトーム、ゲノムの後天的修飾様式の分析（エピゲノミクス）まで最先端の内容だった。

カキゲノムは同じような繰り返しが続く配列が多いのが特徴で、今までゲノム解析は非常に困難

知って得する 日立ITプラットフォームマガジン

変革は、ITから

ビッグデータから未来をつくる“人材”を

ゲノム科学領域におけるビックデータ解析の人材育成で指導的立場にある

ニュース 日経情報ストラテジー

情報処理学会がビッグデータ活用実務フォーラムを第1回を開催

日本経済新聞電子版

2013/10/31

西村 崇＝日経情報ストラテジー （筆者執筆記事一覧）

記事一覧へ >>

5 おすすめ 4 共有 7 ブックマーク 22 28 ツイート

シェア

情報処理学会は2013年10月30日、東京都内でビッグデータを活用する現場担当者を対象にした実務フォーラムを初めて開催した（写真）。

写真●情報処理学会が開催したビッグデータ活用実務フォーラム
[画像のクリックで拡大表示]

ビッグデータを活用しようという動きは、ITの分野にとどまらず、医学や農学など多岐にわたりつつある。代表者である東京農工大学の石井一夫特任教授は「ビッグデータ活用に関心は集まっているものの、実務に携わる人材の育成や実務家同士の情報交換の場が少ない。そこでこのフォーラムを立ち上げた」と話す。

第1回目の実務フォーラムでは、石井特任教授のほか、アマゾンデータ サービス ジャパンの吉荒祐一ソリューションアーキテクトと、ユニバーサル・シェル・プログラミング研究所（USP研究所）の當仲寛哲所長が講演した。

事例　1

判別分析による発現変動遺伝子の組合せの最適化

研究の目的

大きな目的：

麹菌* の遺伝子欠損株を解析し、欠損した遺伝子の機能を明らかにする

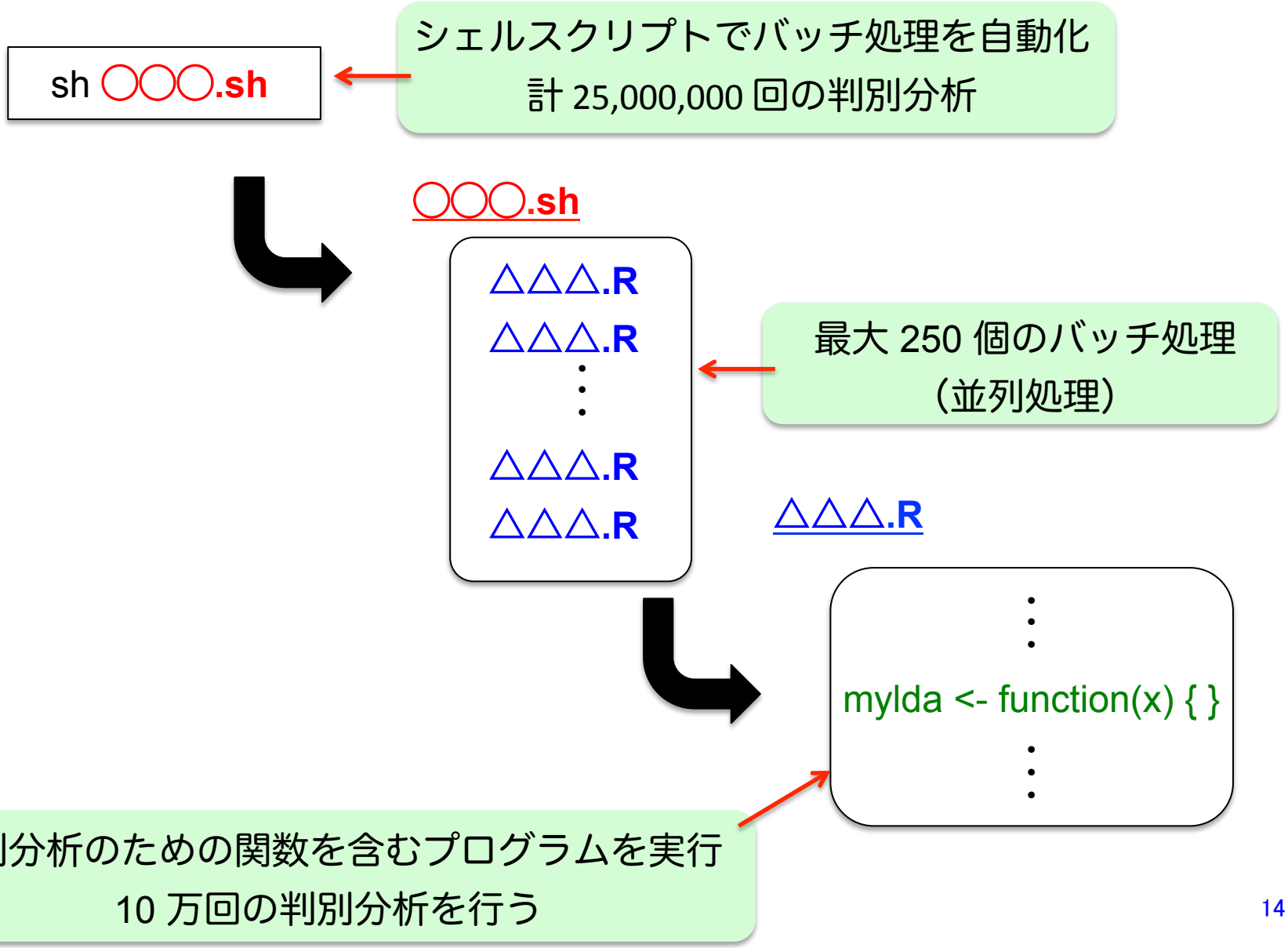
*麹菌：日本酒や醤油、味噌などの日本の伝統的醸造産業に欠かせない真核微生物
日本の “国菌” であると認定されている

解析手法の一つとして、**RNA-Seq 解析**を行った

麹菌のある遺伝子を欠損した株と Control 株の遺伝子の発現量データを取得

遺伝子欠損株で発現が変動した遺伝子を調べることで、
欠損した遺伝子の機能を明らかにするための重要な示唆を与える

解析の流れ


sh ○○○.sh
シェルスクリプトでバッチ処理を自動化
計 25,000,000 回の判別分析
○○○.sh
△△△.R
△△△.R
△△△.R
△△△.R
最大 250 個のバッチ処理
（並列処理）
△△△.R
mylda <- function(x) { }
判別分析のための関数を含むプログラムを実行
10 万回の判別分析を行う

組合せの絞り込み②

Tokyo University of Agriculture and Technology

選抜した Gene 237, 73, 120, 319, 117, 12, 79, 51, 253, 359 それぞれを含む組合せについて、説明変数 5 個の 1 億 2,500 万回の判別分析の結果を Wilks' lambda の小さい順に並べる

Gene 237

遺伝子の組合せ	Wilks' lambda の値
237,185, 26, 148, 306	3.66111675008352e-05
212, 42, 237, 53, 338	6.40126561402781e-05
117, 42, 53, 338, 237	6.72759297517926e-05
191, 237, 12, 3, 55	7.1513427442044e-05
359, 327, 4, 237, 378	7.42064950463597e-05
⋮	⋮

Gene 237 を含む組合せのうち、Wilks' lambda の値が最も小さい組合せを選び、判別分析の結果をプロットすることで評価

（選抜した 10 個の遺伝子について行った）

Student's *t*-test で選んだ組合せの判別分析結果

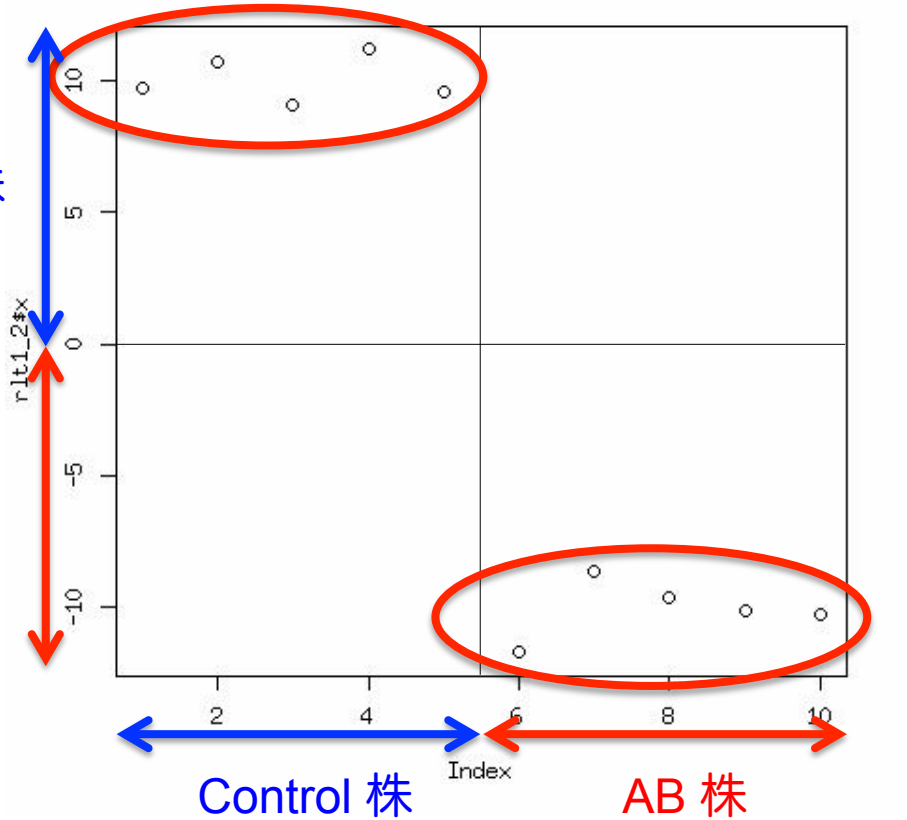
Tokyo University of Agriculture and Technology

Student's *t*-test の *p*-value 低い 5 個の遺伝子の判別分析の結果

判定結果

Control 株

AB 株



実際の株（正解）

単に Student's *t*-test で *p*-value が低い 5 遺伝子を選んだ組合せでも
欠損株とコントロール株を区別することができた

方法：判別関数（数理モデル）の作成

Tokyo University of Agriculture and Technology

```
(rlt1_1 <- lda(as.matrix(data4[,x]), grouping1))
Call:
lda(as.matrix(data4[, x]), grouping = grouping1)

Prior probabilities of groups:
  0   1
0.5 0.5

Group means:
       GeneA      GeneB      GeneC      GeneD      GeneE
0 -0.7755834  0.7996515  0.5910129 -0.8376868  0.6256248
1  0.7755834 -0.7996515 -0.5910129  0.8376868 -0.6256248

Coefficients of linear discriminants:
              LD1
GeneA   139.008572
GeneB   -45.685510
GeneC   -58.397485
GeneD   141.975012
GeneE     5.454758
> apply(rlt1_1$means%*%rlt1_1$scaling,2,mean)
         LD1
-8.526513e-14
> rlt1_2 <- predict(rlt1_1)
> plot(rlt1_2$x)
```

$$y = ax1 + bx2 + cx3 + dx4 + ex5 - f$$

各遺伝子の係数
(= a, b, c, d, e)

定数項 (= f)

まとめ

文部科学省特別経費の支援による「ゲノム科学人材育成プログラム」の中で、
データ分析、統計解析、数理モデリング、プログラミングなどを主流とする統計科学の実践教育を実施。

5年間で述べ２４５名の大学院生に対し、統計科学、情報科学の個別指導、セミナーなどを実施。

指導した大学院生により
100回以上の学会発表、10報以上の投稿論文、
10件以上の受賞などを達成。

課題

大学院での研究の高度化には貢献できたが、
正式カリキュラムではないため、系統だった教育が
困難。

実際にデータ分析が自力でできるようになる
学生は学年に1人か2人というのが実情。

謝辞

東京農工大学　ゲノム人材育成　古崎利紀、石川英明、泉川圭一、高橋信弘
農学研究科　小林拓嗣、山形洋平
その他、農学研究科大学院生、教員の先生方

宇都宮大学　農学研究科大学院生、教員の先生方

茨城大学　農学研究科大学院生、教員の先生方