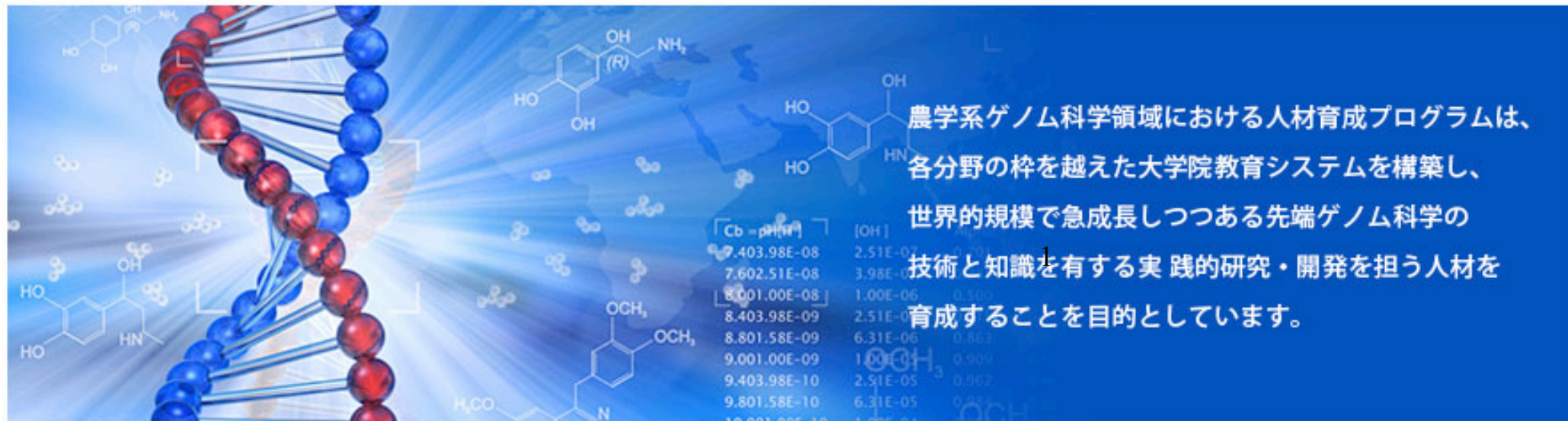


S-PLUS/Rを用いた ゲノム科学のビッグデータ解析

石井一夫 東京農工大学

2013年11月22日
数理システムユースーズカンファレンス2013



農学系ゲノム科学領域における人材育成プログラムは、
各分野の枠を越えた大学院教育システムを構築し、
世界的規模で急成長しつつある先端ゲノム科学の
技術と知識を有する実践的研究・開発を担う人材を
育成することを目的としています。

| [Cb - pH(m)] | [OH] |
|--------------|----------|
| 7.403.98E-08 | 2.51E-07 |
| 7.602.51E-08 | 3.98E-07 |
| 8.001.00E-08 | 1.00E-06 |
| 8.403.98E-09 | 2.51E-06 |
| 8.801.58E-09 | 6.31E-06 |
| 9.001.00E-09 | 1.00E-05 |
| 9.403.98E-10 | 2.51E-05 |
| 9.801.58E-10 | 6.31E-05 |
| 1.00E-09 | 1.00E-04 |

我々のビッグデータ処理の新しい産業応用 広告やゲーム、レコメンだけではない

個別化医療(ライフサイエンス):

精神神経系疾患(うつ病、総合失調症)の
網羅的ゲノム診断法の開発

→ 全人類のゲノム解析と個別化医療実現を目標

ゲノム育種(グリーンサイエンス):

ブルーベリー、オオムギ、イネなどの新品種の開発

環境アセスメント(エコサイエンス):

環境微生物の分布、分類、生態調査

→ 世界遺産に指定された東南アジアの
古代遺跡の環境破壊状況の調査

ゲノム科学におけるビッグデータ分析

本研究室では、大規模データ解析に対し以下の4方法で対応している。
HPCも、モンテカルロも、クラウドも使えるものは何でも使う主義。

1. モンテカルロシミュレーション：大量データから無作為にサンプルを抽出し、元のデータをシミュレーション

2. Big iron (大容量メモリサーバ)による大量並列処理 (HPC)
HP社の協力により、4TBメモリ、CPU: Xeon E7 (80 コア、160スレッド)の大容量メモリ解析サーバを使用

3. Hadoop による分散処理システム (クラウドを利用) (後日報告)
Amazon Elastic MapReduce (Amazon EMR) プラットフォーム

4. Hadoop によらない分散処理システム
シェルスクリプトベースの分散処理。usp-BOA (USP研究所) を利用。大量データのクオリティチェック (後日報告)

本日は、上記の1,2,の方法の適用例を紹介する。

次世代シーケンサーの例

イルミナ社 GAIIX, MiSeq,
ロシュダイアグノスティクス社 454
ライフテクノロジーズ社 IonProton
など



GAIIX



MiSeq

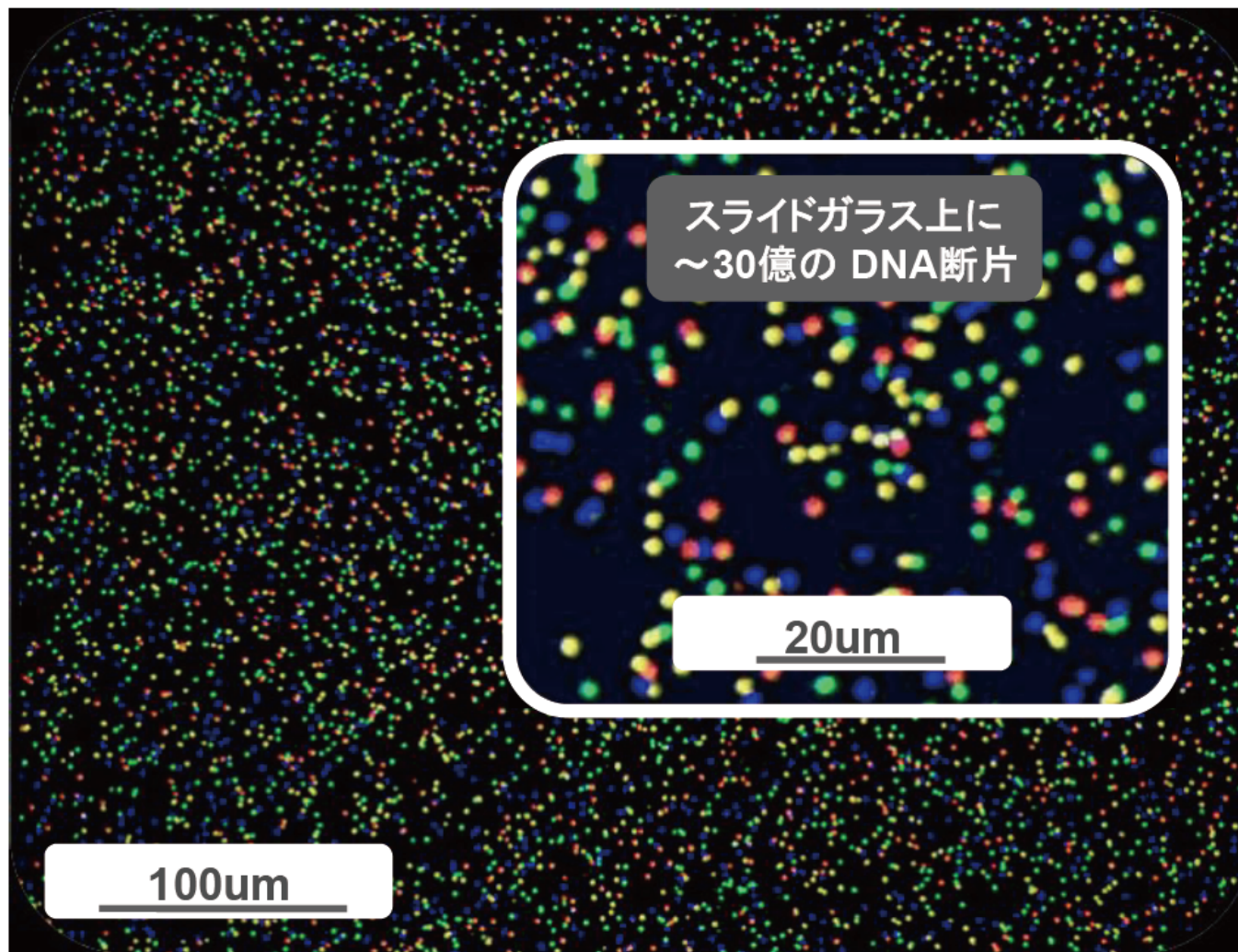
次世代シーケンサー

固相基板上に、DNA断片を固相化し、これを蛍光色素+酵素反応などを用いて、同時並列的に解読。CCD カメラで撮影+コンピュータで処理。

一度に、約数十～数千塩基のDNA断片を数十万から数億のDNA断片同時に解読できます。

数日から、数時間でヒトゲノム DNAを30億塩基解読することも可能。

次世代シーケンサーのDNA解読画像

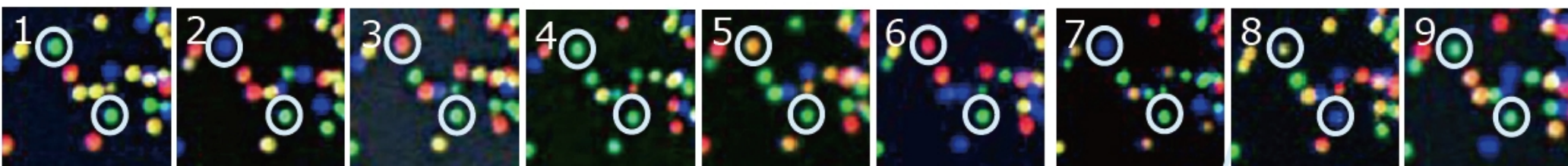


イルミナ社資料より

次世代シーケンサーのDNA解読画像

色のイメージからDNAの配列を決定

T



T

イルミナ社資料より

次世代シーケンサーのデータ解析

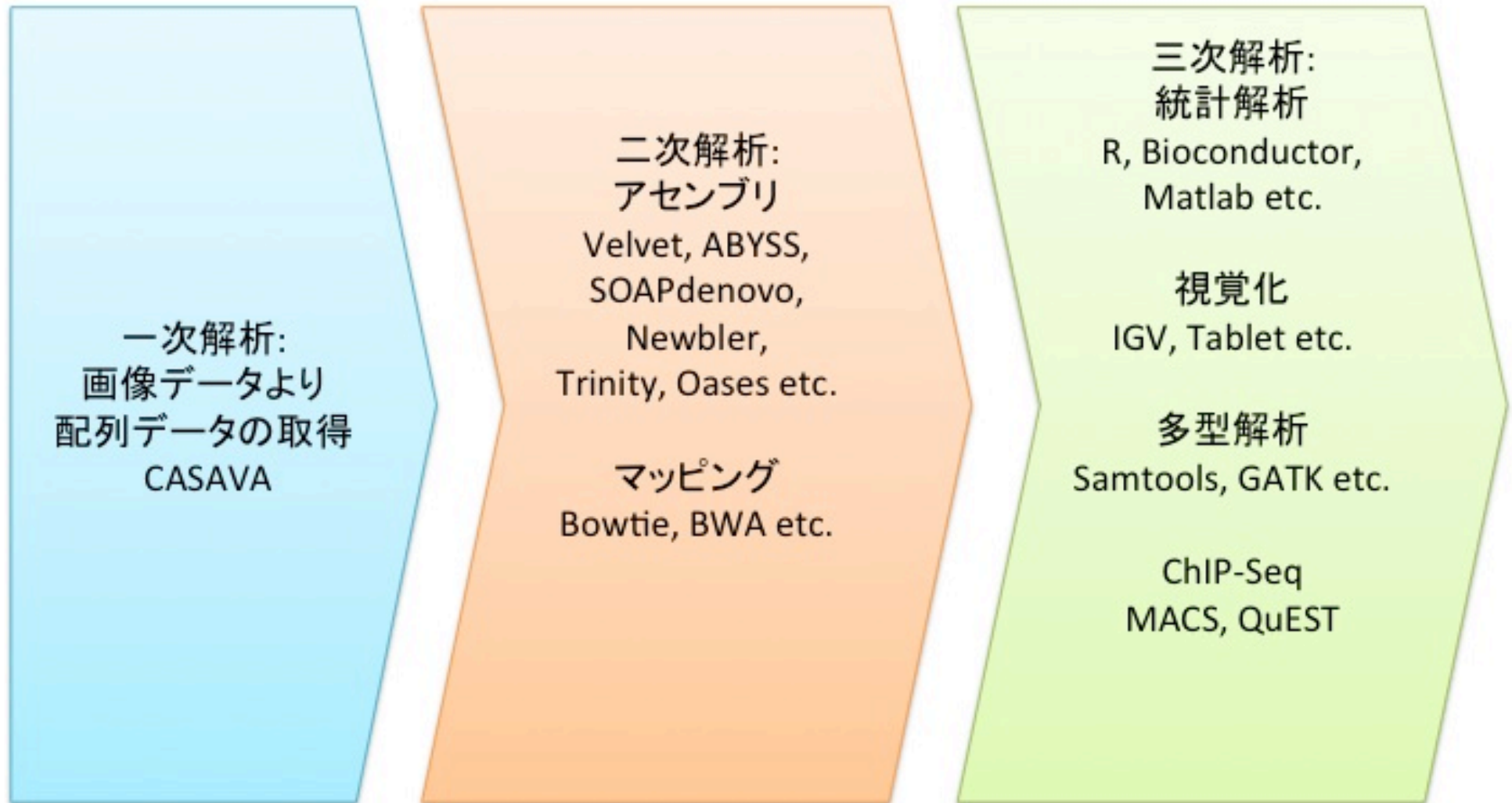
データ解析システムは主に、Linux をベースとしたフリーソフトウェアを用いて構築。

テキスト処理と数値計算を駆使し、awk, grep, cut, cat, sedなどのコマンドと、Perl/Python/Ruby などのスクリプト言語、S-PLUS/R、Octave(Matlab) などの統計解析ソフトを活用する。

専用解析ソフトも多く開発、C、C++、Java など、開発。

データ解析は3段階

Conventional Genomic Analysis System on Linux Server



使用するソフトウェア群

3段階のデータ解析

1段階 画像処理 → DNA塩基配列取得

**2段階 DNA塩基配列を連結、整列、編集
アセンブリ 配列を相互に連結
マッピング 参照配列に整列**

**3段階 統計処理、視覚化、
データマイニング**

解析ソフト群

1段階 画像処理 → DNA塩基配列取得
CASAVA

2段階 配列を集計、編集
アセンブリ 配列を相互に連結
Velvet、SOAPdenovo、Trinity、
マッピング 参照配列に整列
Bowtie、BWA、、

3段階 統計処理、視覚化、
データマイニング **S-PLUS/R,**
¹¹BLAST、、

Hadoopを用いた 次世代シーケンサーのデータ解析

コンピュータクラスタや、クラウドの活用。

Hadoop MapReduce などのビッグデータフレームワークの活用例も増えてきている。

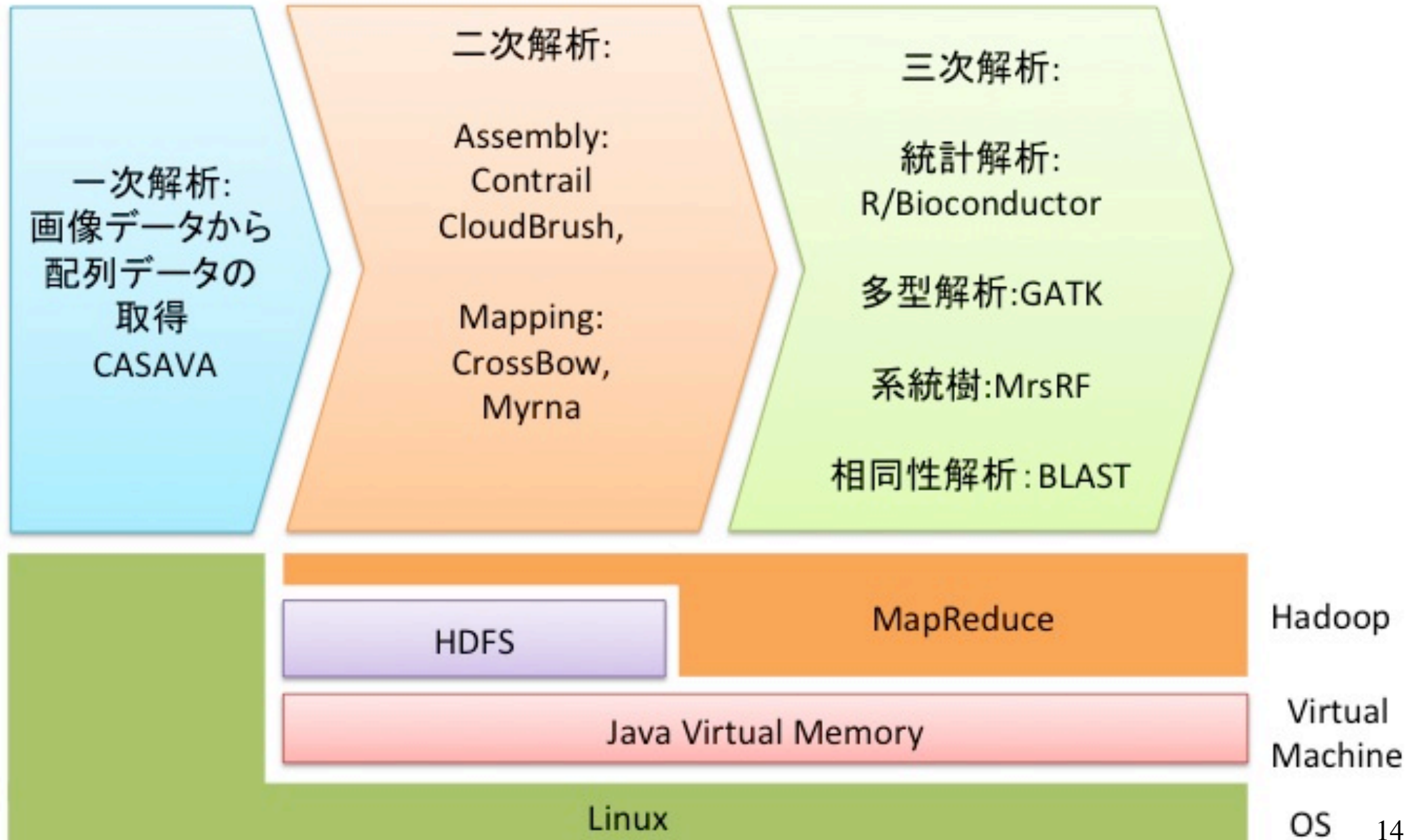
なぜ クラウド??

個人レベル、1研究室レベルで、コンピュータクラスタを構築することは、コスト的に無理。

クラウドでしか、コンピュータクラスタを使えない。

Hadoop を個人レベルで構築するひとつの方法として、クラウドサービスの利用を検討。

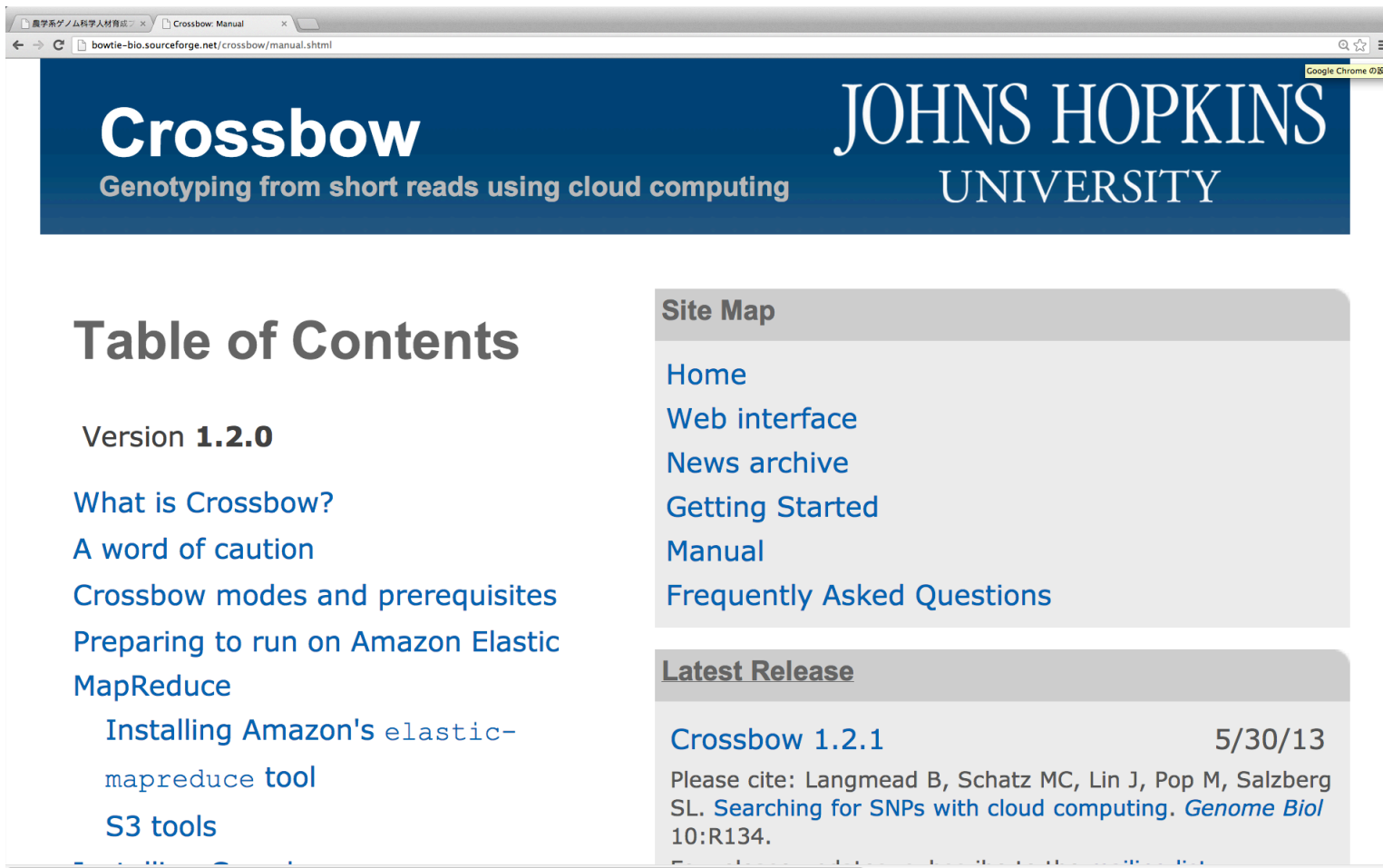
Hadoop 上で動作する解析システム



使用するソフトウェア群

Crossbow

Hadoopを使用して次世代シーケンサーデータをマッピングし、多型を検出するソフトウェア。



The screenshot shows a web browser displaying the Crossbow website. The browser's address bar shows the URL `bowtie-bio.sourceforge.net/crossbow/manual.shtml`. The website has a dark blue header with the "Crossbow" logo and the tagline "Genotyping from short reads using cloud computing". To the right of the header is the "JOHNS HOPKINS UNIVERSITY" logo. The main content area is divided into three sections: "Table of Contents", "Site Map", and "Latest Release".

Table of Contents

- Version **1.2.0**
- [What is Crossbow?](#)
- [A word of caution](#)
- [Crossbow modes and prerequisites](#)
- [Preparing to run on Amazon Elastic MapReduce](#)
- [Installing Amazon's elastic-mapreduce tool](#)
- [S3 tools](#)

Site Map

- [Home](#)
- [Web interface](#)
- [News archive](#)
- [Getting Started](#)
- [Manual](#)
- [Frequently Asked Questions](#)

Latest Release

Crossbow 1.2.1 5/30/13


Please cite: Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL. [Searching for SNPs with cloud computing](#). *Genome Biol* 10:R134.

AWSでのCrossbowの利用

農学系ゲノム科学人材育成プロジェクト

AMI: Crossbow short read

← → ↺ <https://aws.amazon.com/amis/crossbow-short-read-genotyping-i386> 🔍 ☆ ☰



Sign Up My Account


AWS Products & Solutions ▾

AMIs from All Sources ▾ 🔍


Developer Resources

Developer Resources

- Amazon Machine Images (AMIs)
- Articles & Tutorials
- Customer Apps
- Developer Tools
- Documentation
- Public Data Sets
- Release Notes
- Sample Code & Libraries

 AMI: Crossbow short read genotyping: i386

Amazon Machine Images (AMIs) > AMI Details

 **Community Contributed Software**

- Amazon Web Services provides links to these packages as a convenience for customers. AWS has not been reviewed or screened by AWS.
- Please review this software to ensure it meets your needs before using it.

Base AMI for running Crossbow for short-read genotyping tasks. This AMI was built by running the "create-image" script with HADOOP_VERSION=0.20.0, ARCH='i386', and version 1.6.0_16

Details

Submitted by:


Launch

GATK (Genome Analysis Toolkit)

Javaベースで動作するMapreduceのフレームワークを取り入れた遺伝子多型解析用ソフトウェア

Intro to the GATK

www.broadinstitute.org/gatk/about/index



[Home](#) [About](#) [Guide](#) [Community](#) [Downloads](#)

Search

About

[Introduction to the GATK](#)

[Citing the GATK](#)

[GATK in print](#)

[User stories](#)

[Who we are](#)

Introduction to the GATK

What is the GATK?

[Using the GATK](#)

[Typical Workflows](#)

[High Performance](#)

[Getting Help](#)

[Licensing & Source Code](#)

What is the GATK?

Simply what it says on the can: a Toolkit for Genome Analysis

Say you have ten exomes and you want to identify the rare mutations they all have in common – the GATK can do that. Or you need to know which mutations are specific to a group of patients, as opposed to a healthy cohort – the GATK can do that too. In fact, the GATK is the industry standard for such analyses.

But wait, there's more!

Because of the way it is built, the GATK is highly generic and can be applied to all kinds of datasets and genome analysis problems. It can be used for discovery as well as for validation. It's just as happy handling exomes as whole genomes. It can use data generated with a variety of different sequencing technologies. And although it was originally developed for human genetics, the GATK has evolved to handle genome data from any organism, with any level of ploidy. Your plant has six copies of each chromosome? Bring it on.



The toolkit provides a wide set of tools that can be chained into workflows, taking advantage of the common architecture and powerful engine.




The GATK can handle a variety of organism genomes in addition to humans.

So what's in the can?

At the heart of the GATK is an industrial-strength infrastructure and engine that handle data access, conversion and traversal, as well as high-performance computing features. On top of that lives a rich ecosystem of specialized tools, called "walkers", that you can use out of the box, individually or chained into scripted workflows, to perform anything from simple data diagnostics to complex "reads-to-results" analyses.

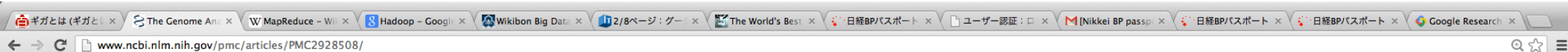
Some typical workflows are detailed on the next page of this section. Please see the [Technical Documentation](#) section for a complete list of tools and their capabilities.

High Performance →



© Broad Institute 2012

GATK の論文



The Genome Analysis Toolkit: A MapReduce framework for analyzing n...

Genome Res. 2010 September; 20(9): 1297–1303.



Aa

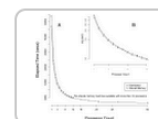
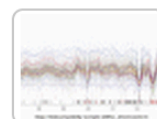
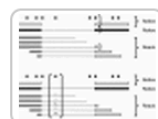


The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data

Aaron McKenna, Matthew Hanna, [...], and Mark A. DePristo

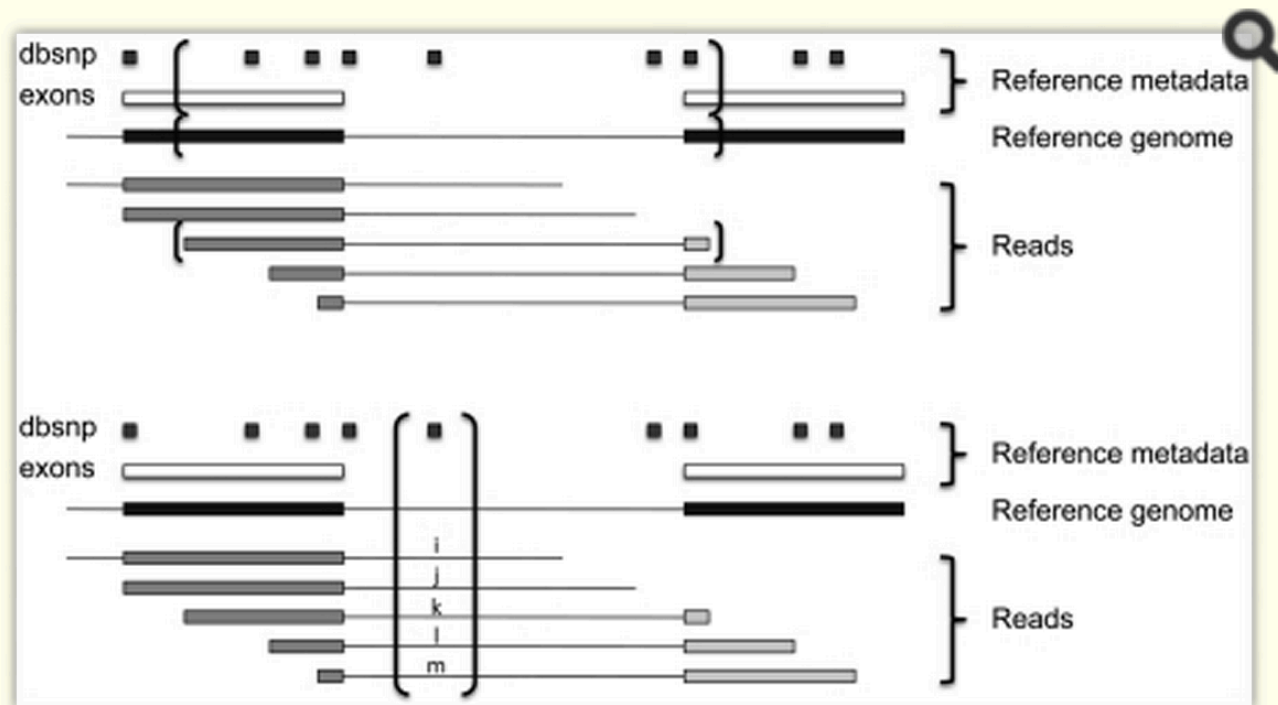
[Additional article information](#)

Abstract



GATKによるSNPコーリング (GATK の論文より)

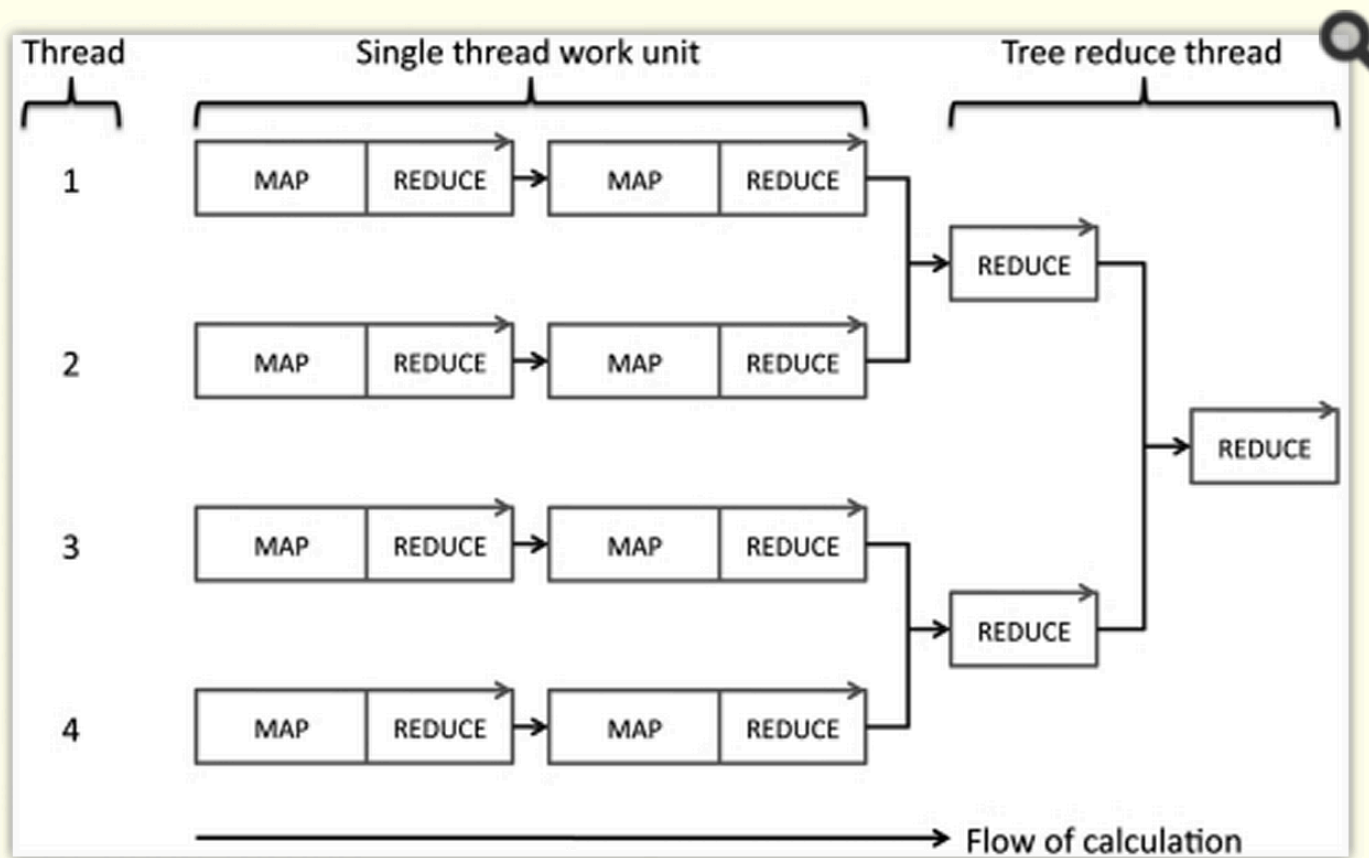
Figure 1.



Read-based and locus-based traversals. Read-based traversals provide a

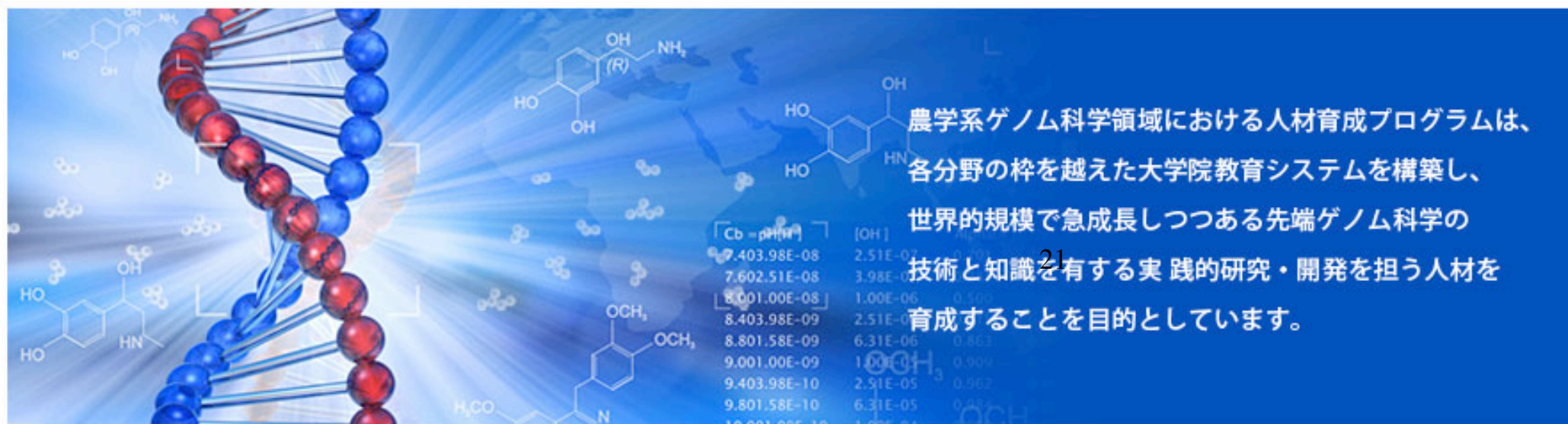
MapReduceをモデルにしたGATKのプロセス(GATK の論文より)

Figure 2.



解析例 I

S-PLUS/Rを用いた 次世代シーケンサーデータの品質管理



農学系ゲノム科学領域における人材育成プログラムは、
各分野の枠を越えた大学院教育システムを構築し、
世界的規模で急成長しつつある先端ゲノム科学の
技術と知識を有する実践的研究・開発を担う人材を
育成することを目的としています。

| [Cb - pH(m)] | [OH] |
|--------------|----------|
| 7.403.98E-08 | 2.51E-07 |
| 7.602.51E-08 | 3.98E-07 |
| 8.001.00E-08 | 1.00E-06 |
| 8.403.98E-09 | 2.51E-06 |
| 8.801.58E-09 | 6.31E-06 |
| 9.001.00E-09 | 1.00E-05 |
| 9.403.98E-10 | 2.51E-05 |
| 9.801.58E-10 | 6.31E-05 |
| 1.00E-09 | 1.00E-04 |

次世代シーケンサーデータの品質管理

次世代シーケンサーデータのクオリティ
チェックに関する問題点。

イルミナ社の場合

フローセル上に DNA 反応クラスタを作らせる。

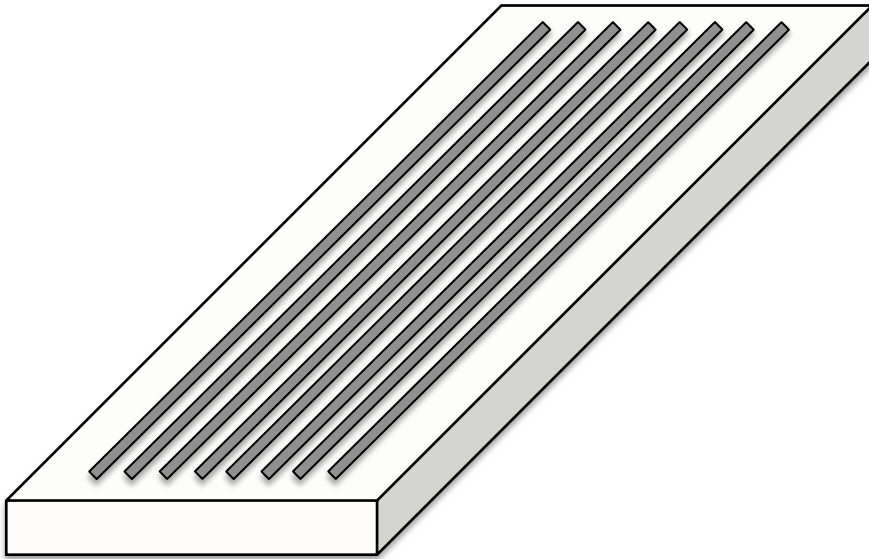
(1) サンプル濃度の間違い、(2) 試薬濃度の間違い、(3) 操作の荒さ、(4) 電圧の不適、(5) データ転送のコマ落ち

様々な原因により、クオリティ悪化が発生。

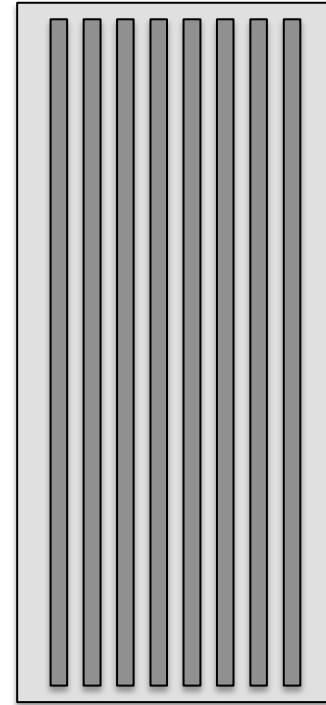
次世代シーケンサーデータの品質管理

イルミナ社の場合

フローセル上に DNA 反応クラスタを作らせる。



フローセルの図
(斜め上から見た図)



フローセルの図
(上から見た図)

次世代シーケンサーデータの品質管理

ランニングコストが高いために、詳細なチェックを行ない、クオリティのいいものだけを使用する工夫が必要。

数十～数百塩基のDNA 断片を数十万～数億断片のクオリティチェック

→ 計算機的にかなり高い負荷

効率のよいチェックを行ない悪品質データの除去を行なうことが必要

→ データのサンプリングなどの工夫。

次世代シーケンサーから出力される配列データ

●fastqファイル

@HWI-ST977:153:D0J0KACXX:4:1101:1636:1851 1:N:0:TAGCTT
NGGTCCGGCTTTGAACCCCTGACAGGAAGGTATTATGCTGATCACGATG
CAACATGACAGATCGGCTCATGAAGCTTGGACTTGCTGTTCTCCTCTTTA
CG

+

#4=BDFFFHHHHJJJJJHJJGGIIFGIIJJJJJJJJJJJJJJJJJJGIJHGHEFF
FDDCCEC@ACACDDDDCCCDDAACCCDDCACDCDCC?

@HWI-ST977:153:D0J0KACXX:4:1101:1705:1877 1:N:0:TAGCTT

NTATCTTGACAGATTTTCTAGACTCATCCCAAGTTCTTGACCTAGCGCTG
ACAGAATTTGCTAAAATATGCTTATTCCGGTGCCAACTCCGTGGTATGCC

A

+

#1=DFFFFHHHHHJJJJJJJJJJJJJJJIIEGHIHGGIJJJIJJJJIIIGJJJIJGI
JJJJIJJIFHEFHGFFDDEEDDDDDDDDDDDDDDDDDDD

4行で1組

配列名

DNA配列

記号

クオリティデー
タ

4行1組の配列データが並び、全体で数百万～数億行に及ぶ

従来のクオリティチェック(FASTQC)

FASTQ ファイルのPhred quality score
をもとに評価

Phred score = $-10 * \log_{10}(\text{error probability})$

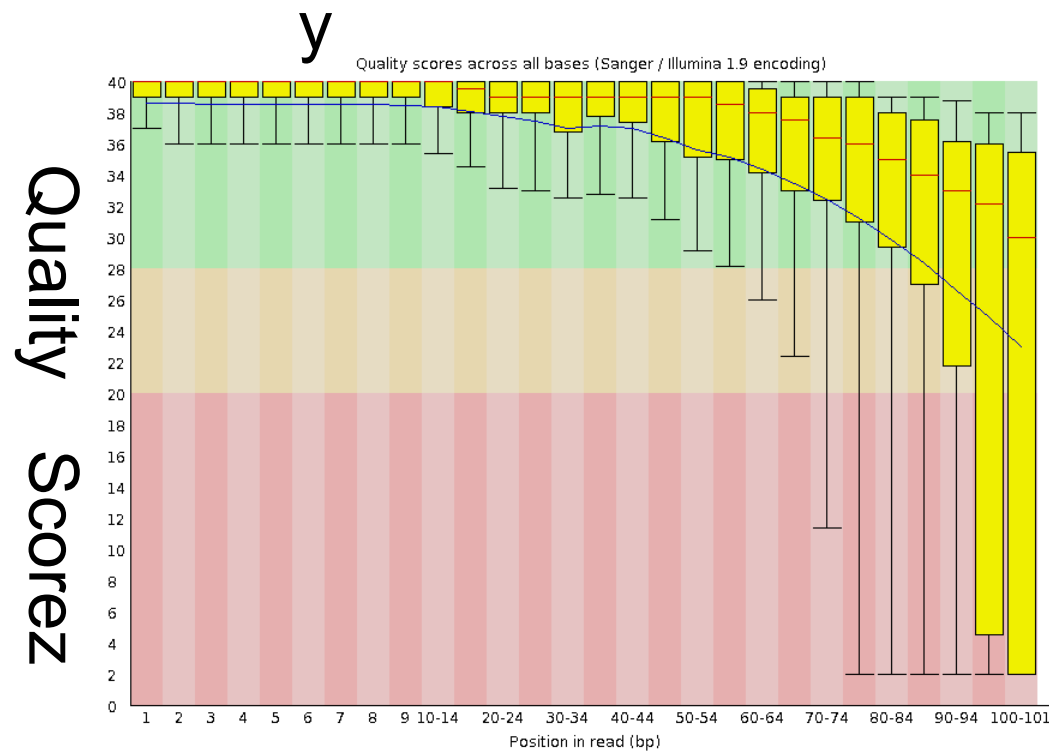
0~40段階：40が最良、0が最悪

40字のアルファベット + 数字 + 記号で表現

最初の20万リードの平均値でデータを評価

配列データのクオリティチェック

FastQC → 配列データのクオリティチェックを行うフリーソフトウェア
per_base_quality



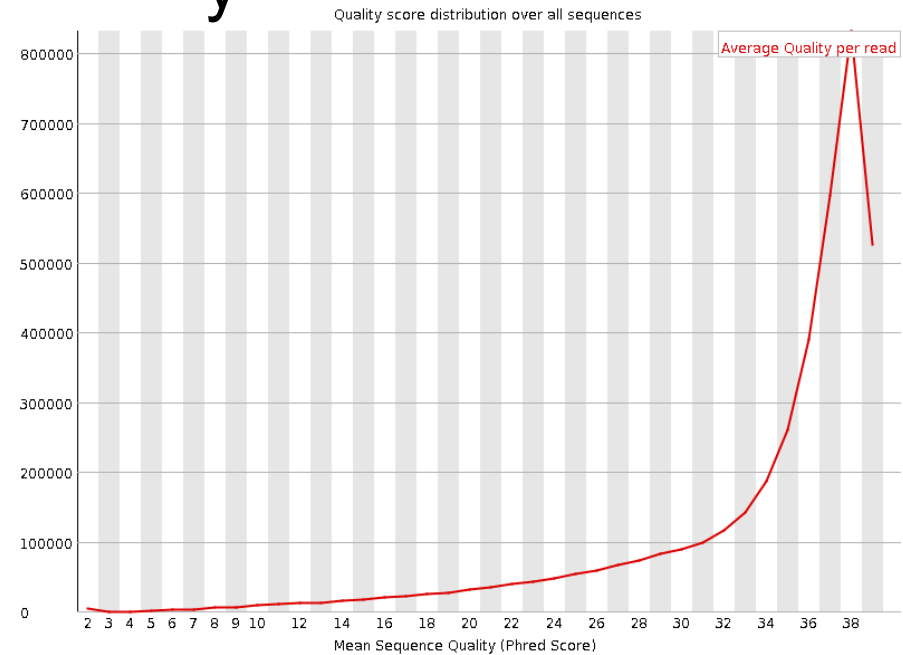
1文字目



DNA配列

頻度

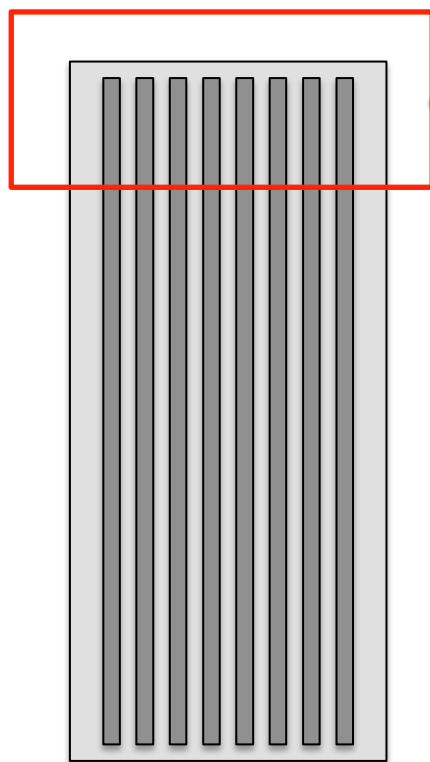
per_base_quality



Quality Score

クオリティチェックの問題点

GAIIxの場合には、フローセル上でDNA反応クラスタを作る



**FastQCでは、
最初の20万個
をチェックして
いる。**

(1) サンプル濃度の間違い

(2) 試薬濃度の間違い

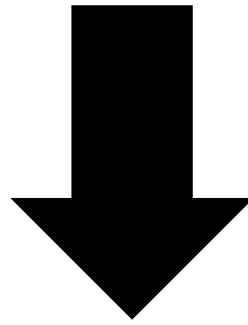
(3) 操作の荒さ

などの原因によりクオリティ
が悪化することがある。

**フローセルの図
(上から見た図)**

研究目的

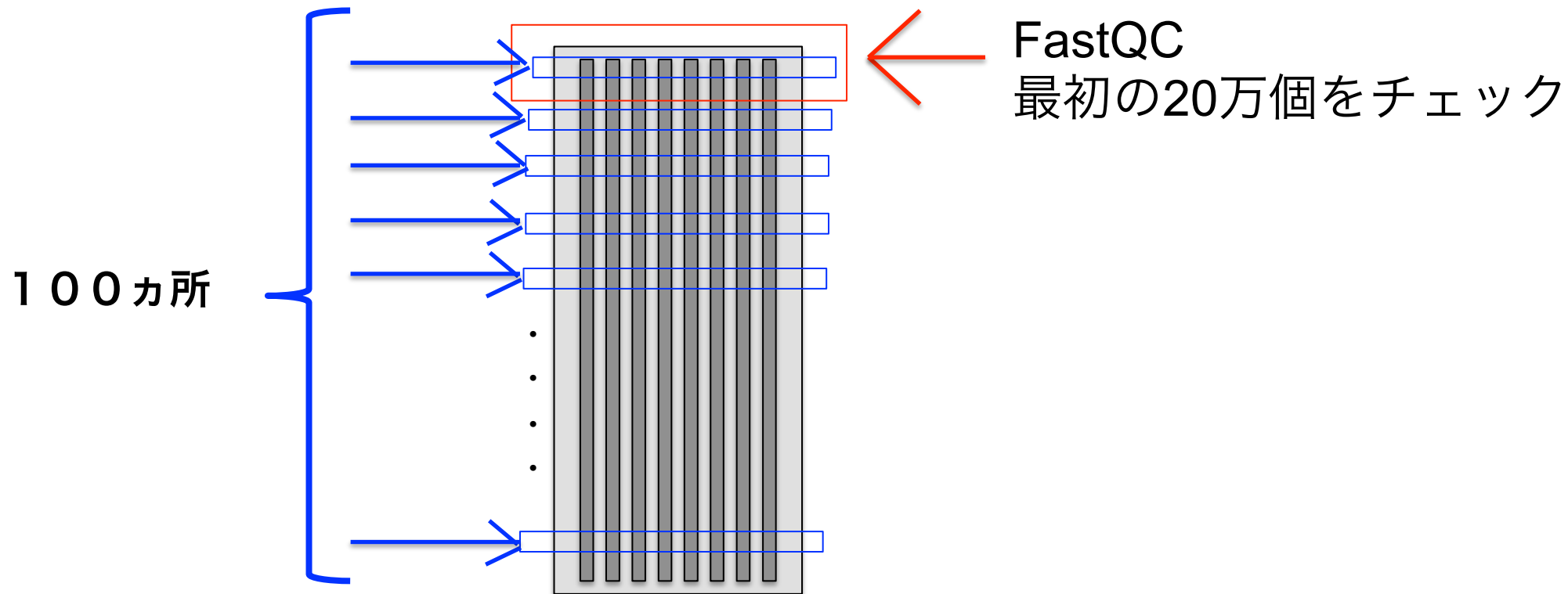
FastQCでは各ファイルの最初の20万個のクオリティデータを使用して計算を行う



配列データが数千万個に及ぶ場合には、クオリティデータ全体を正確に把握することが困難

**ファイル全体のクオリティデータを評価する
方法確立する**

モンテカルロ法によるクオリティチェック①

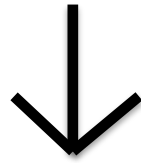


モンテカルロ法

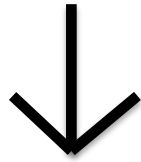
全体を100等分し、100カ所から1000個ずつリードを抽出

モンテカルロ法によるクオリティチェック②

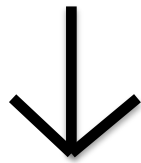
統計解析ソフト「**S-PLUS/R**」を使用し乱数を発生



乱数に基づき、fastqファイルから無作為にクオリティデータを抽出



クオリティデータを数値データに変換

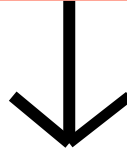


S-PLUS/Rの関数hist ()、density ()、heatmap () などを使用して全体のクオリティを評価

S-PLUS/Rによる乱数の発生

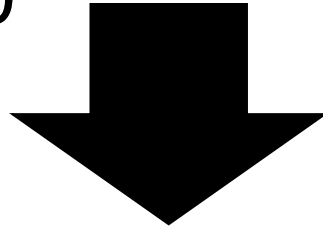
使用サンプル：L197 (Read1：15057415 reads、Read2：
15057415 reads)

15057415read



100区間に等分

1 ～ 150574、150575～311148、・・・、
14906872～15057400



各区間において、**S-PLUS/R**の関数runif()を使用して、それぞれ1000個の乱数を発生させる（**一様乱数**）。

さらに、発生させた100区間分の乱数のデータをcbind()によつての結合し、100行×1000列の乱数表としてまとめた。

●Fastqデータ

4列で1組

リード名

DNA配列

記号

クオリティデー
タ

➡ 作製した乱数表の数字に当たる行のクオリティデータを抽出

100区間で1000個のクオリティデータを、無作為に抽出

クオリティデータの数値変換

無作為に抽出したデータをperlのスク립トを使用して数値データに変換

クオリティデータ

#4=BDFFFGHHHHHJJIIJJIHJJGGIIFGIIIJJIIJJJJJJJJJJJJJJJJJJGIJHGHEFFFDD
CCEC@ACACDDDDCCCDDAACCCDDCACDCDCC?

Table 1 ASCII Characters Encoding Q-scores 0–40

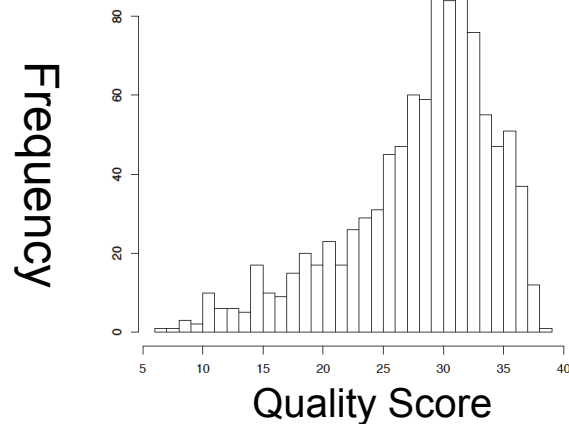
| Symbol | ASCII Code | Q-Score | Symbol | ASCII Code | Q-Score | Symbol | ASCII Code | Q-Score |
|--------|------------|---------|--------|------------|---------|--------|------------|---------|
| ! | 33 | 0 | / | 47 | 14 | = | 61 | 28 |
| " | 34 | 1 | 0 | 48 | 15 | > | 62 | 29 |
| # | 35 | 2 | 1 | 49 | 16 | ? | 63 | 30 |
| \$ | 36 | 3 | 2 | 50 | 17 | @ | 64 | 31 |
| % | 37 | 4 | 3 | 51 | 18 | A | 65 | 32 |
| & | 38 | 5 | 4 | 52 | 19 | B | 66 | 33 |
| ' | 39 | 6 | 5 | 53 | 20 | C | 67 | 34 |
| (| 40 | 7 | 6 | 54 | 21 | D | 68 | 35 |
|) | 41 | 8 | 7 | 55 | 22 | E | 69 | 36 |
| * | 42 | 9 | 8 | 56 | 23 | F | 70 | 37 |
| + | 43 | 10 | 9 | 57 | 24 | G | 71 | 38 |
| , | 44 | 11 | : | 58 | 25 | H | 72 | 39 |
| - | 45 | 12 | ; | 59 | 26 | I | 73 | 40 |
| . | 46 | 13 | < | 60 | 27 | | | |

得られたクオリティの数値データから各リードのクオリティの平均値を計算し、それを基にヒストグラム、密度推定、ヒートマップを作製し、クオリティデータを評価した。

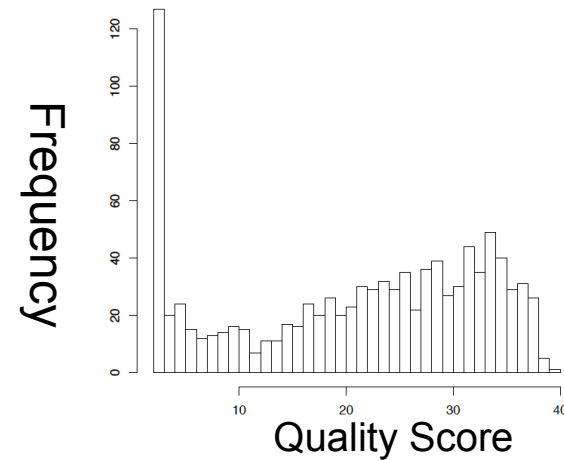
ヒストグラムによる評価

Rの関数hist()を使用して各区間でのヒストグラムを作製した。

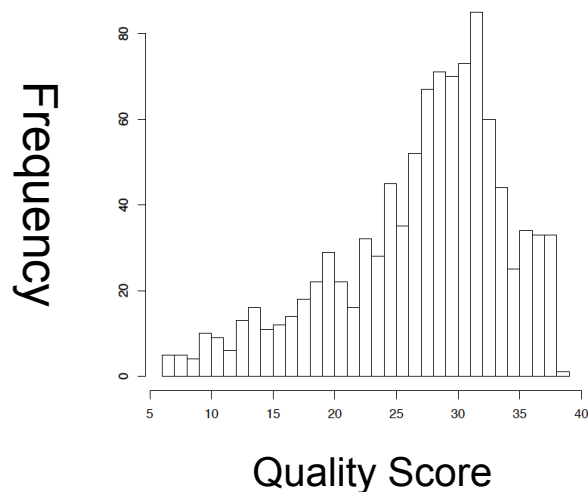
R1の1番目の区間（1～150574）のクオリティ



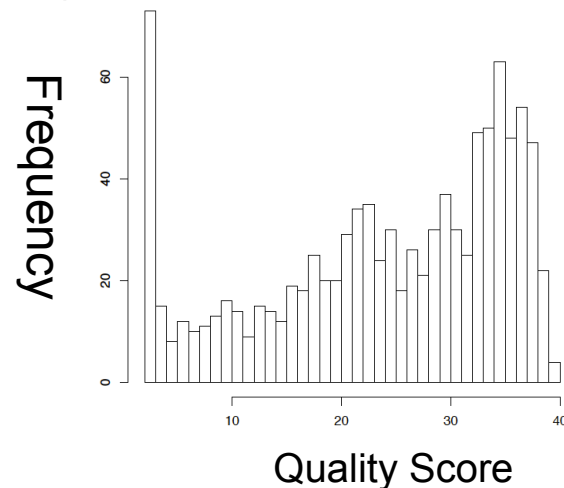
R2の1番目の区間（1～150574）のクオリティ



R1の41番目の区間（12045921～12196494）のクオリティ



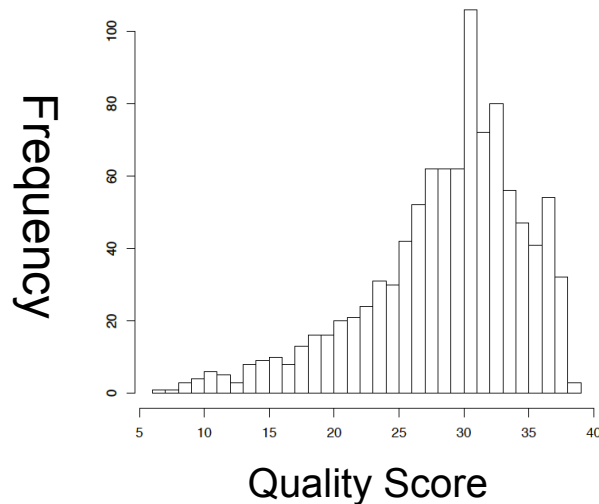
R2の81番目の区間（12045921～12196494）のクオリティ



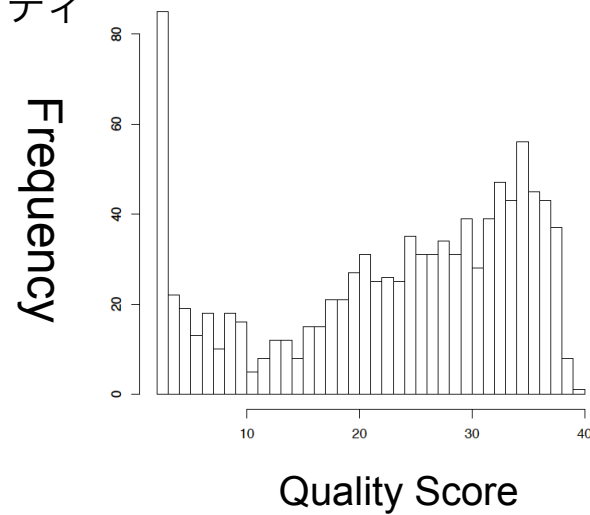
密度推定による評価

Rの関数hist()及びdensity()を使用して各区間でのクオリティチェックを行った。

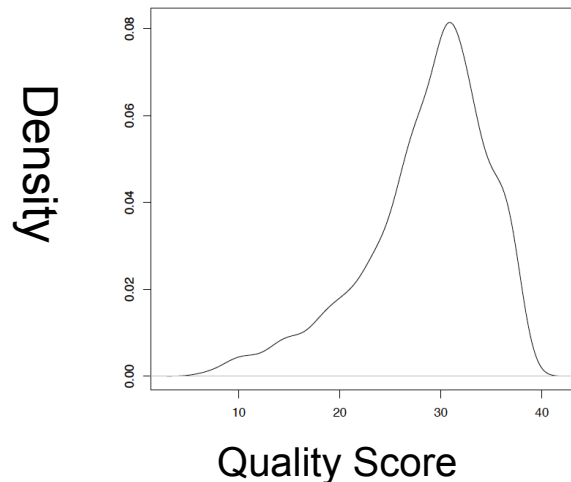
R1の41番目の区間（12045921 ～12196494 ）のクオリティ



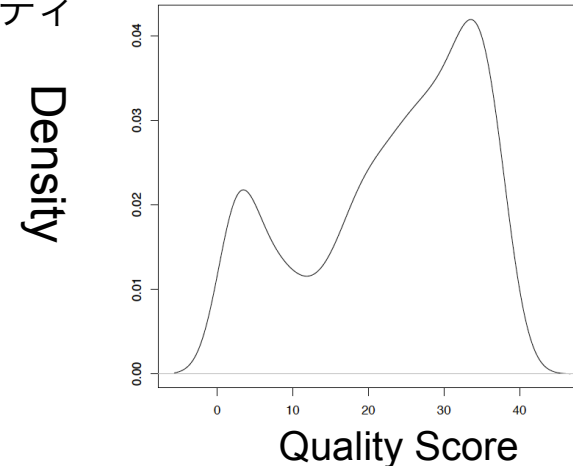
R2の81番目の区間（12045921 ～12196494 ）のクオリティ



R1の41番目の区間（12045921 ～12196494 ）のクオリティ



R2の81番目の区間（12045921 ～12196494 ）のクオリティ

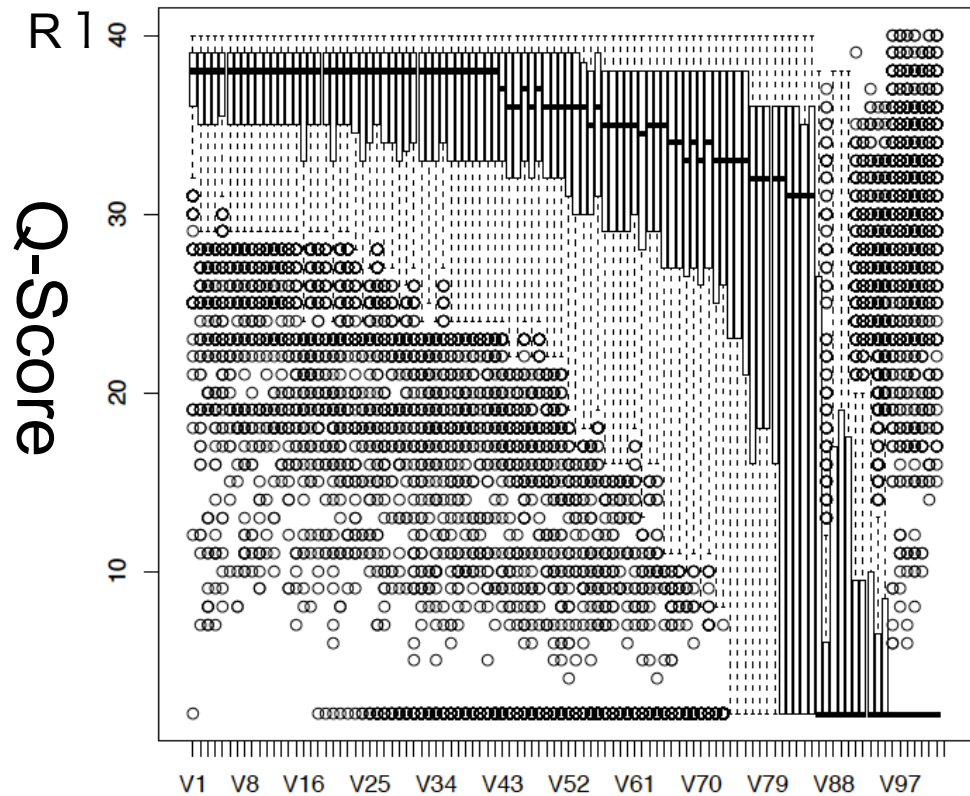


箱ひげ図による評価

Rの関数boxplot()を使用して各区間での各塩基長ごとの
を行った。

DNA配列

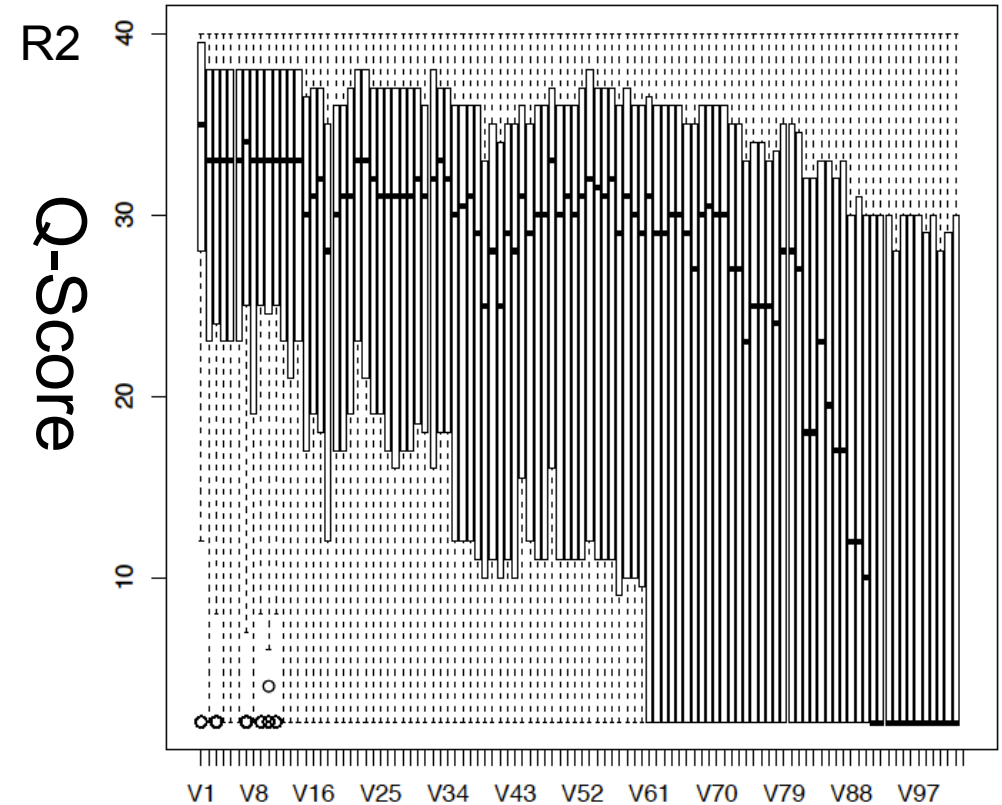
NTATCTTGACAGATTTTCTAGACTCATCCCAAGTTCTTGACCTAGCGCTGAC
AGAATTTGCTAAAATATGCTTATTCCGGTGCCAACTCCGTGGTATGCCA



1文字目



DNA配列



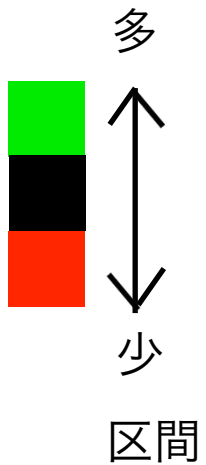
1文字目



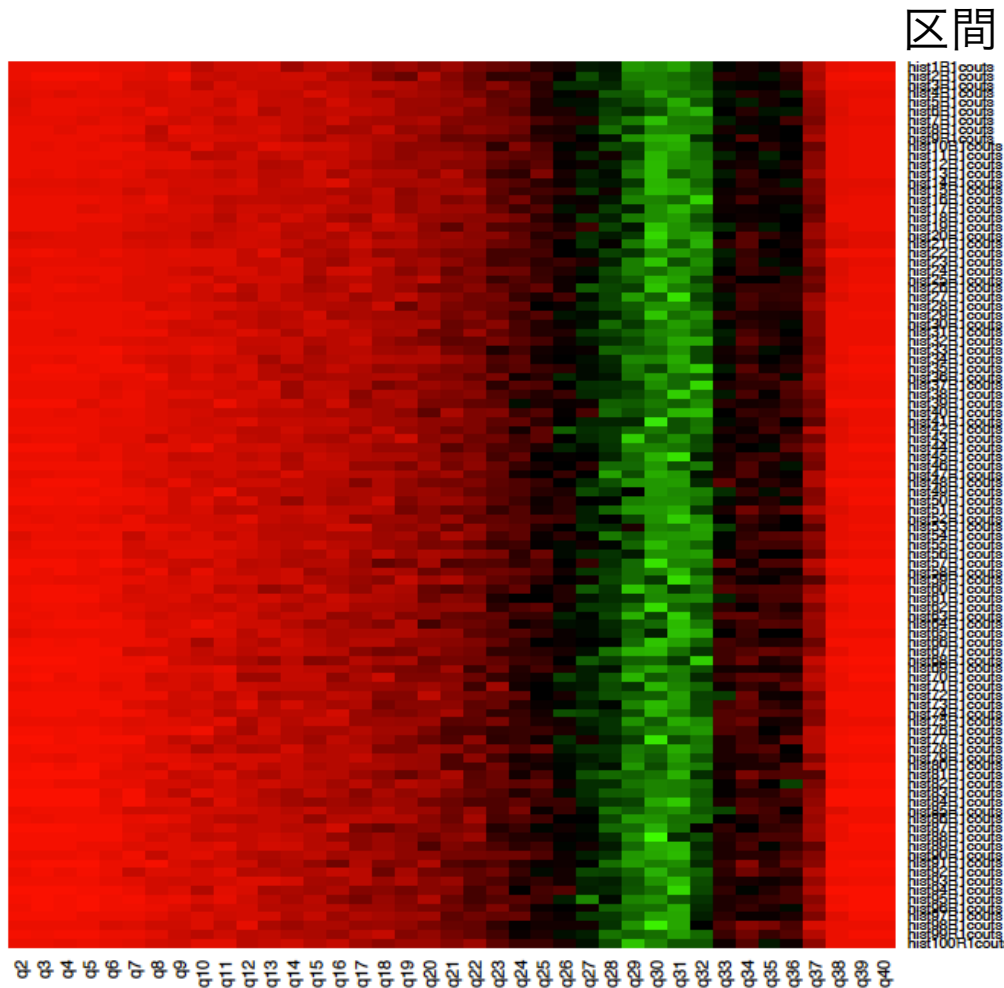
DNA配列

ヒートマップによる評価

heatmap()関数を使用して全体のヒートマップを作製した。

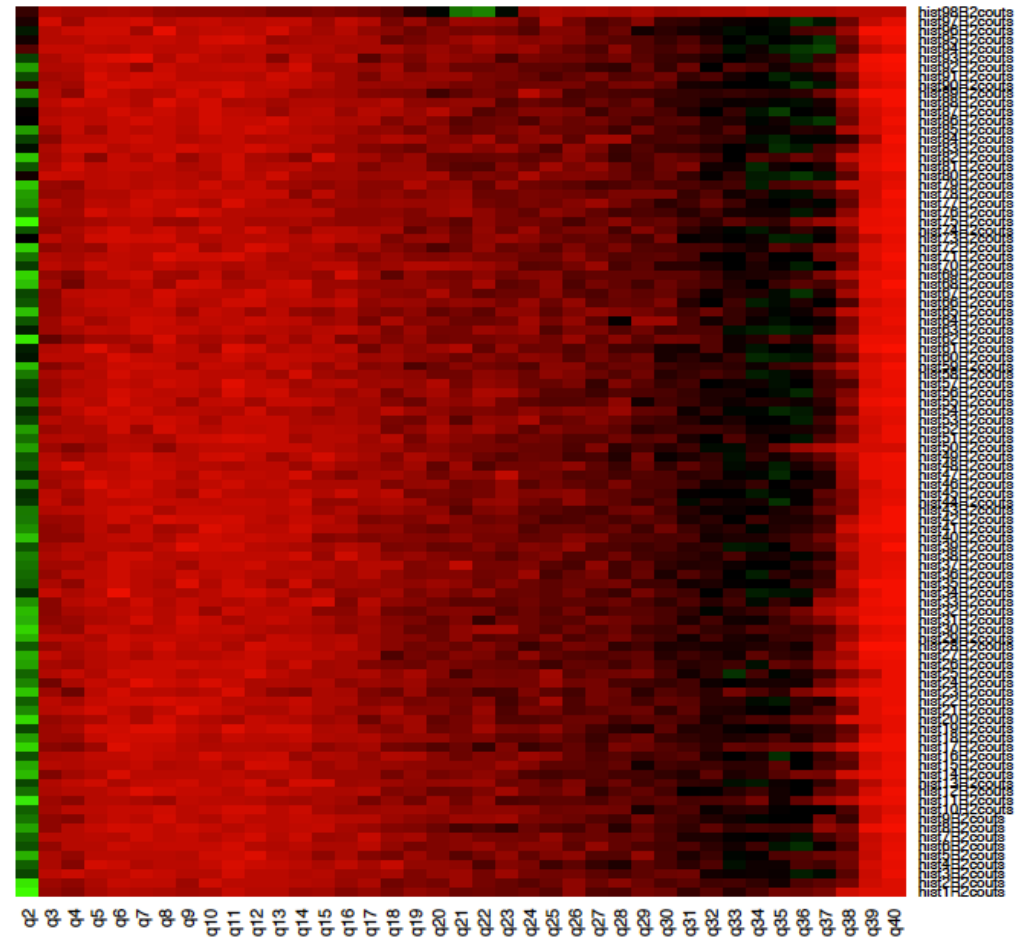


R1のクオリティデータ



Quality Score

R2のクオリティデータ



Quality Score

まとめ

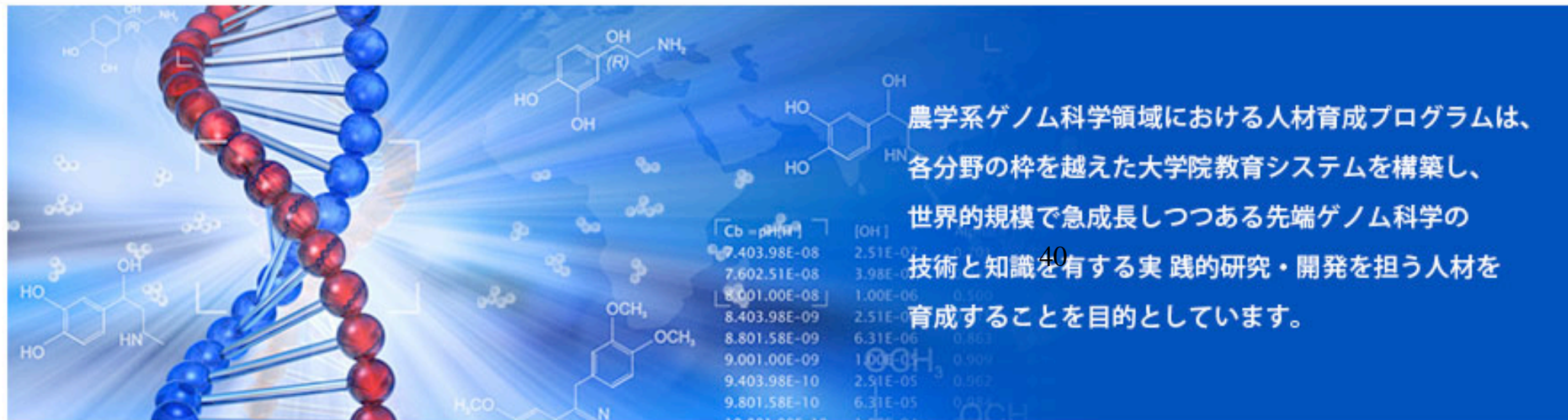
1) 次世代シーケンサーデータのクオリティは均一でないケースがある。特に、ランが失敗した例では不均一性顕著に出る。

2) クオリティチェックに使用されているソフト (FASTQC) は最初 20 万塩基しかチェックしていないので、ランが失敗した場合のクオリティチェックデータの解釈は注意が必要。

→FASTQCの結果 = ラン全体のクオリティ
になるとは限らない。

解析例 II

S-PLUS/Rを用いた分子進化の 分岐図作成の最適化



目的:

447種類の遺伝子を用いた28種の近縁生物種の分岐図を作成する

従来、分岐図(進化系統樹)は、ミトコンドリアや16S RNAなど比較的短い配列情報をもとに作成していた。

近年の、ゲノム解析技術の進歩から、ゲノム全体に渡る配列情報を用いて分岐図を作成することが可能となった。

その遺伝子(配列)の組合せはほぼ無限にあり、どのような配列をもとに分岐図を作成するかは、議論の余地がある。

しかし、この検討には大量の計算が必要である。

今回、モンテカルロ法による無作為抽出と、ビッグデータ処理による最適化法を確立することを目的とした。

方法： 分岐図作成の最適化検討のためのシナリオ

整列遺
伝子

1、447種類の遺伝子のマルチプルアラインメントデータを使用。

組合せ生成

2、447種類の遺伝子の組合せ数を計算。組合せのパターンを生成。
447種類の遺伝子のうち4つまでの遺伝子を除く組合せの数が、**1656133808**であつたので、今回はこれで、最適化法の確立の検討を行なった。447個の遺伝子の総当たりの組合せは、 $6.109568e+99$ であり、これを検討することは現実には無理。

モンテカル
口法

3、**1656133808**の数から、一様乱数を発生。該当する遺伝子の組合せたマルチプルアラインメントを作成(**ブートストラップ法**)。

分岐図作成

4、分岐図を作成し、分岐パターンを抽出。最尤法による推定により分岐図を作成(分岐図作成ソフトRAxMLを使用。)

分岐パター
ン選択

5、分岐パターンを集計して、最適化分岐図を作成し、分岐図作成に対し寄与度の高い遺伝子を検出した。

実施結果：

447個の遺伝子から任意の個数の遺伝子を選択する組合せ数の計算 (with **S-PLUS/R**)

```
> choose(447,1) # 447個の遺伝子から1遺伝子を選択
```

```
[1] 447
```

```
> choose(447,1)+choose(447,2) # 2遺伝子を選択
```

```
[1] 100128
```

```
> choose(447,1)+choose(447,2)+choose(447,3)
```

```
[1] 14886143 # 3遺伝子を選択
```

```
>choose(447,1)+choose(447,2)+choose(447,3)+  
choose(447,4) # 4遺伝子を選択
```

```
[1] 1656133808 <- 今回はこれで検証(該当する遺伝子を除く)
```

```
>choose(447,1)+choose(447,2)+choose(447,3)+.....+
```

```
choose(447,447) # 447遺伝子を選択
```

```
[1] 6.109568e+99
```

実施結果：

組合せ数列を生成するためのPerlのスクリプト

```
#!/usr/local/bin/perl
use strict;
use warnings;
our $number = 447;
our $test_flag = 0;

reflex("", 1, $number) if $test_flag;
for(my $rest = 1; $rest <= $number; $rest++){
    reflex("", 1, $rest);
}
sub reflex{
    my($computed, $start, $rest) = @_;
    if($rest == 1){
        for(my $i = $start; $i <= $number; $i++){
            print "$computed$i¥n";
        }
        return;
    }else{
        $rest--;
        my $end = $number - $rest;
        for(my $i = $start; $i <= $end; $i++){
            reflex("$computed$i,", $i + 1, $rest);
        }
    }
}
```

447個の遺伝子から
4個までを選択する
1656133808個の
組合せ数列を産生

実施結果：

乱数の発生と対応する組合せ数列の選択

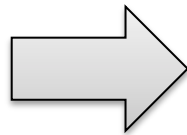
1、**S-PLUS/R**を用いて一様乱数(runif())を発生)

```
run1 <- round(runif(100000, min = 1, max = 1656133808)) # 10万個の乱数を発生
```

2、乱数に対応する組合わせ数列を選択

発生させた乱数

81571247
365913044
1023396239
644416555
691641727
1268663549
982640498
670232037
516837044
938086069



乱数に対応する数列

5,121,150,372
27,29,176,229
95,160,302,413
51,199,232,430
56,126,142,201
135,342,359,446
89,303,405,437
54,98,149,213
39,234,259,417
84,108,135,403

3、447個のうち乱数に対応する遺伝子を除き、残りのすべての遺伝子を連結させたマルチプルアラインメントファイルを作成

実施結果： RAxML の実行による分岐図の作成

マルチプルアラインメントデータのfasta 形式ファイルから **phylip 形式ファイルへの変換**

```
/usr/local/EMBOSS-6.4.0/bin/seqret fasta::1.fasta phylip::1r.phy
```

分岐図作成ソフト**RAxMLの実行**

```
/usr/local/packages/RAxML/RAxML-7.2.8-ALPHA/raxmlHPC-PTHREADS -f a -x 12345  
-p 12345 -# 20 -m GTRGAMMA -s 1.phy -n 1phy.out -T 16
```

RAxML の出力結果から以下のコマンドにより、数値や記号などを取り去り、
分岐図の簡素化パターンに変換

```
less RAxML_bestTree.200.out | tr -d [0-9] | tr -d ":" | tr -d "." | tr -d ";"  
> RAxML_bestTree.200SUMMARY.out
```

実施結果： RAxML の出力結果を分岐図パターンに変換

RAxML の出力結果から以下のコマンドにより、数値や記号などを取り去り、分岐図パターンに変換

```
less RAxML_bestTree.out | tr -d [0-9] | tr -d ":" | tr -d "." | tr -d ";" > RAxML_bestTreeSUMMARY.out
```

==> RAxML_bestTree.out <== # RAxML の出力結果

```
((x:0.00531404813722460411,b:0.00582680929801783314):0.02597126625043840939,(m:0.03645324955994154459,((i:0.02130122160079299734,g:0.03144790765745542060):0.00424151281867054565,u:0.02948686769656536782):0.02040360046940021752):0.01521774994899611003):0.00840806219060770237,(((z:0.34344767189954156228,(t:0.14262613954560879326,l:0.09611024306570406517):0.12482727188754781655):0.17738679389459199864,(((q:0.09220903636333667441,c:0.07849740402964605623):0.02162213710044687265,((((e:0.02209622018870339294,k:0.02893283119099681472):0.00897154535047478552,p:0.08526167426794276083):0.01393193949473354662,(f:0.01318356630444799359,(n:0.01952490523202302791,&:0.01693127308779425119):0.00813604928815400003):0.07817988855466992404):0.02263432315084732208,(r:0.04590445767519640841,h:0.04637315388999797144):0.03865318679664145329):0.00789191373814705256,(v:0.03284611917996409919,d:0.02250210568318532570):0.17923032798411692168):0.01001049487098192720):0.10325734189742048763,(y:0.20086249354886381857,(j:0.18321752975378832740,#:0.20967985390326032702):0.12992858329809614526):0.01880523845322217696):0.03857414591748902638):0.06620101031901222399,(o:0.03854648520093560682,s:0.03552704927452698946):0.12706855170728659221):0.05630830036902149949,w:0.13933629626394547496):0.07157780409461377003,a:0.03782021720159066402):0.0;
```

==> RAxML_bestTreeSUMMARY.out <== # 変換された簡素化分岐図パターン

```
((x,b),(m,((i,g),u))),((((z,(t,l)),((q,c),((((e,k),p),(f,(n,&))),r,h)),v,d)),y,(j,#))),o,s),w),a)
```

実施結果： 分岐図パターンの集計

分岐図パターンの頻度を計数することで数値化

以下のように、先行20回の試行解析のうち

19回: (((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((((e,k),p),(f,(n,&))), (r,h)), (v,d))), (y,(j,#)))), (o,s)),w),a)

1回: (((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((v,d),((((e,k),p),(f,(n,&))), (r,h))))), (y,(j,#)))), (o,s)),w),a)

となり最適化分岐図が求められた。（ただし、この程度では、検討に足りない、。）

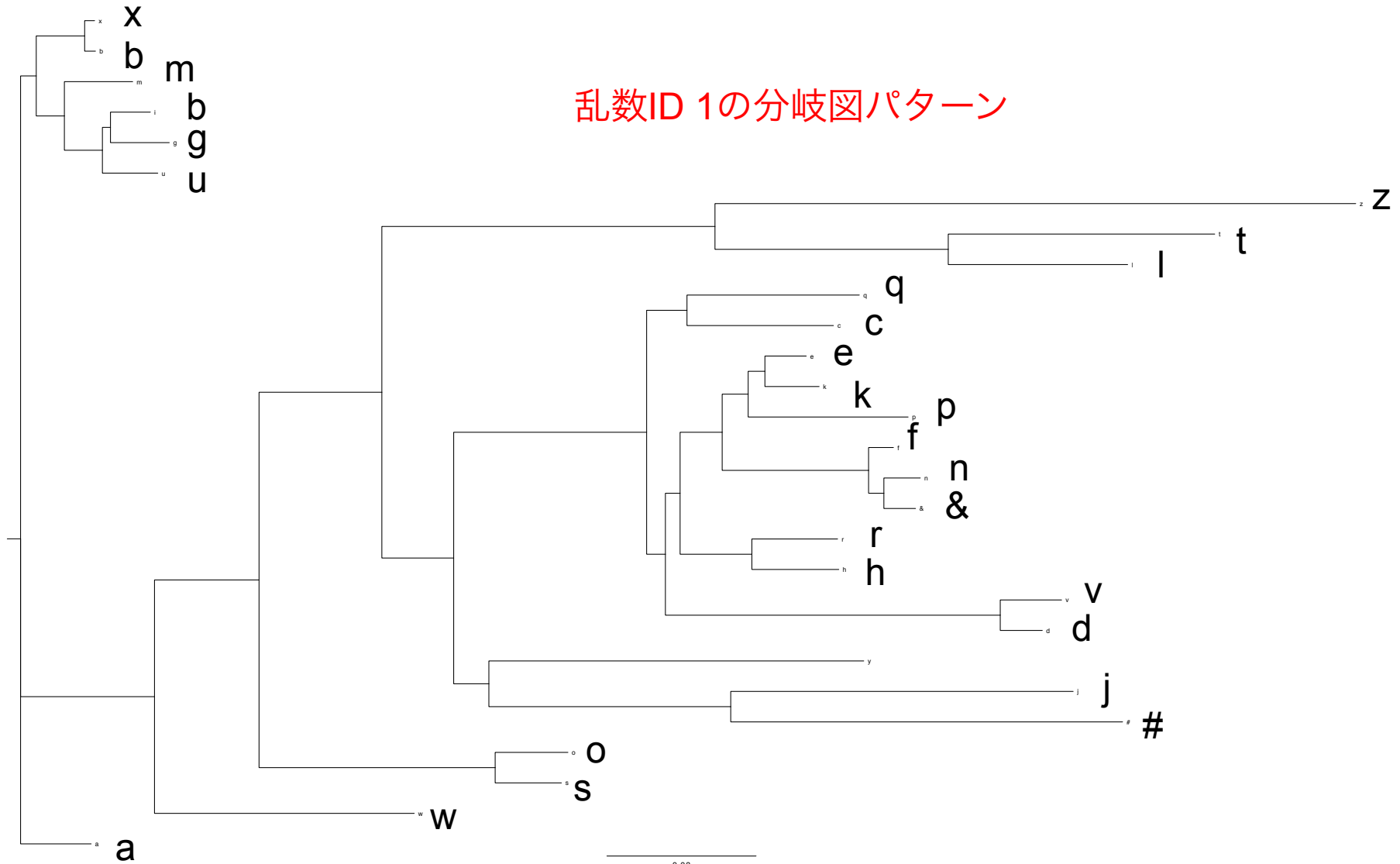
また、189、203、337、393が分岐図作成に強く寄与していることが推定された。

| 乱数ID | 遺伝子1 | 遺伝子2 | 遺伝子3 | 遺伝子4 | Process Name | Tree Pattern | | | | | | |
|------|------|------|------|------|--------------|---|--|--|--|--|--|--|
| 1 | 5 | 121 | 150 | 372 | xaaN1 | (((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((((e,k),p),(f,(n,&))), (r,h)), (v,d))), (y,(j,#)))), (o,s)),w),a) | | | | | | |
| 2 | 27 | 29 | 176 | 229 | xaaN2 | (((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((((e,k),p),(f,(n,&))), (r,h)), (v,d))), (y,(j,#)))), (o,s)),w),a) | | | | | | |
| 3 | 95 | 160 | 302 | 413 | xaaN3 | (((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((((e,k),p),(f,(n,&))), (r,h)), (v,d))), (y,(j,#)))), (o,s)),w),a) | | | | | | |
| 4 | 51 | 199 | 232 | 430 | xaaN4 | (((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((((e,k),p),(f,(n,&))), (r,h)), (v,d))), (y,(j,#)))), (o,s)),w),a) | | | | | | |
| 5 | 56 | 126 | 142 | 201 | xaaN5 | (((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((((e,k),p),(f,(n,&))), (r,h)), (v,d))), (y,(j,#)))), (o,s)),w),a) | | | | | | |
| 6 | 135 | 342 | 359 | 446 | xaaN6 | (((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((((e,k),p),(f,(n,&))), (r,h)), (v,d))), (y,(j,#)))), (o,s)),w),a) | | | | | | |
| 7 | 89 | 303 | 405 | 437 | xaaN7 | (((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((((e,k),p),(f,(n,&))), (r,h)), (v,d))), (y,(j,#)))), (o,s)),w),a) | | | | | | |
| 8 | 54 | 98 | 149 | 213 | xaaN8 | (((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((((e,k),p),(f,(n,&))), (r,h)), (v,d))), (y,(j,#)))), (o,s)),w),a) | | | | | | |
| 9 | 39 | 234 | 259 | 417 | xaaN9 | (((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((((e,k),p),(f,(n,&))), (r,h)), (v,d))), (y,(j,#)))), (o,s)),w),a) | | | | | | |
| 10 | 84 | 108 | 135 | 403 | xaaN10 | (((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((((e,k),p),(f,(n,&))), (r,h)), (v,d))), (y,(j,#)))), (o,s)),w),a) | | | | | | |
| 11 | 107 | 277 | 308 | 420 | xaaN11 | (((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((((e,k),p),(f,(n,&))), (r,h)), (v,d))), (y,(j,#)))), (o,s)),w),a) | | | | | | |
| 12 | 132 | 379 | 399 | 432 | xaaN12 | (((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((((e,k),p),(f,(n,&))), (r,h)), (v,d))), (y,(j,#)))), (o,s)),w),a) | | | | | | |
| 13 | 20 | 223 | 367 | 414 | xaaN13 | (((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((((e,k),p),(f,(n,&))), (r,h)), (v,d))), (y,(j,#)))), (o,s)),w),a) | | | | | | |
| 14 | 23 | 96 | 240 | 421 | xaaN14 | (((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((((e,k),p),(f,(n,&))), (r,h)), (v,d))), (y,(j,#)))), (o,s)),w),a) | | | | | | |
| 17 | 20 | 243 | 277 | 295 | xaaN17 | (((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((((e,k),p),(f,(n,&))), (r,h)), (v,d))), (y,(j,#)))), (o,s)),w),a) | | | | | | |
| 18 | 6 | 219 | 248 | 317 | xaaN18 | (((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((((e,k),p),(f,(n,&))), (r,h)), (v,d))), (y,(j,#)))), (o,s)),w),a) | | | | | | |
| 20 | 56 | 181 | 237 | 313 | xaaN20 | (((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((((e,k),p),(f,(n,&))), (r,h)), (v,d))), (y,(j,#)))), (o,s)),w),a) | | | | | | |
| 21 | 220 | 247 | 327 | 390 | xaaN21 | (((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((((e,k),p),(f,(n,&))), (r,h)), (v,d))), (y,(j,#)))), (o,s)),w),a) | | | | | | |
| 22 | 189 | 203 | 337 | 393 | xaaN22 | (((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((v,d),((((e,k),p),(f,(n,&))), (r,h))))), (y,(j,#)))), (o,s)),w),a) | | | | | | |
| 23 | 149 | 251 | 300 | 316 | xaaN23 | (((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((((e,k),p),(f,(n,&))), (r,h)), (v,d))), (y,(j,#)))), (o,s)),w),a) | | | | | | |
| 24 | 142 | 205 | 277 | 284 | | | | | | | | |

実施結果： 最適化された分岐図パターン

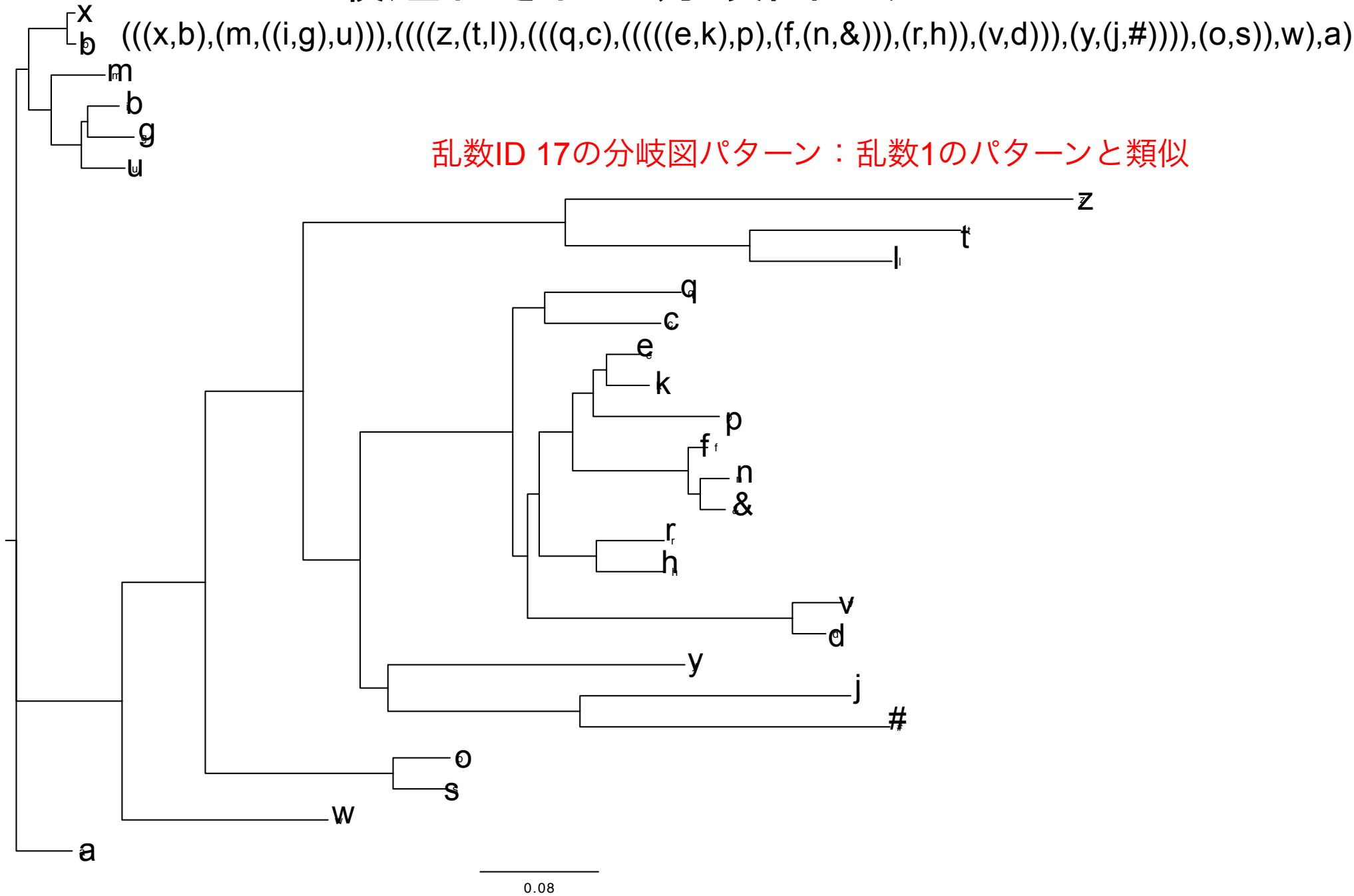
((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((((e,k),p),(f,(n,&))),r,h)),v,d))),y,(j,#))),o,s),w),a)

乱数ID 1の分岐図パターン



実施結果： 最適化された分岐図パターン

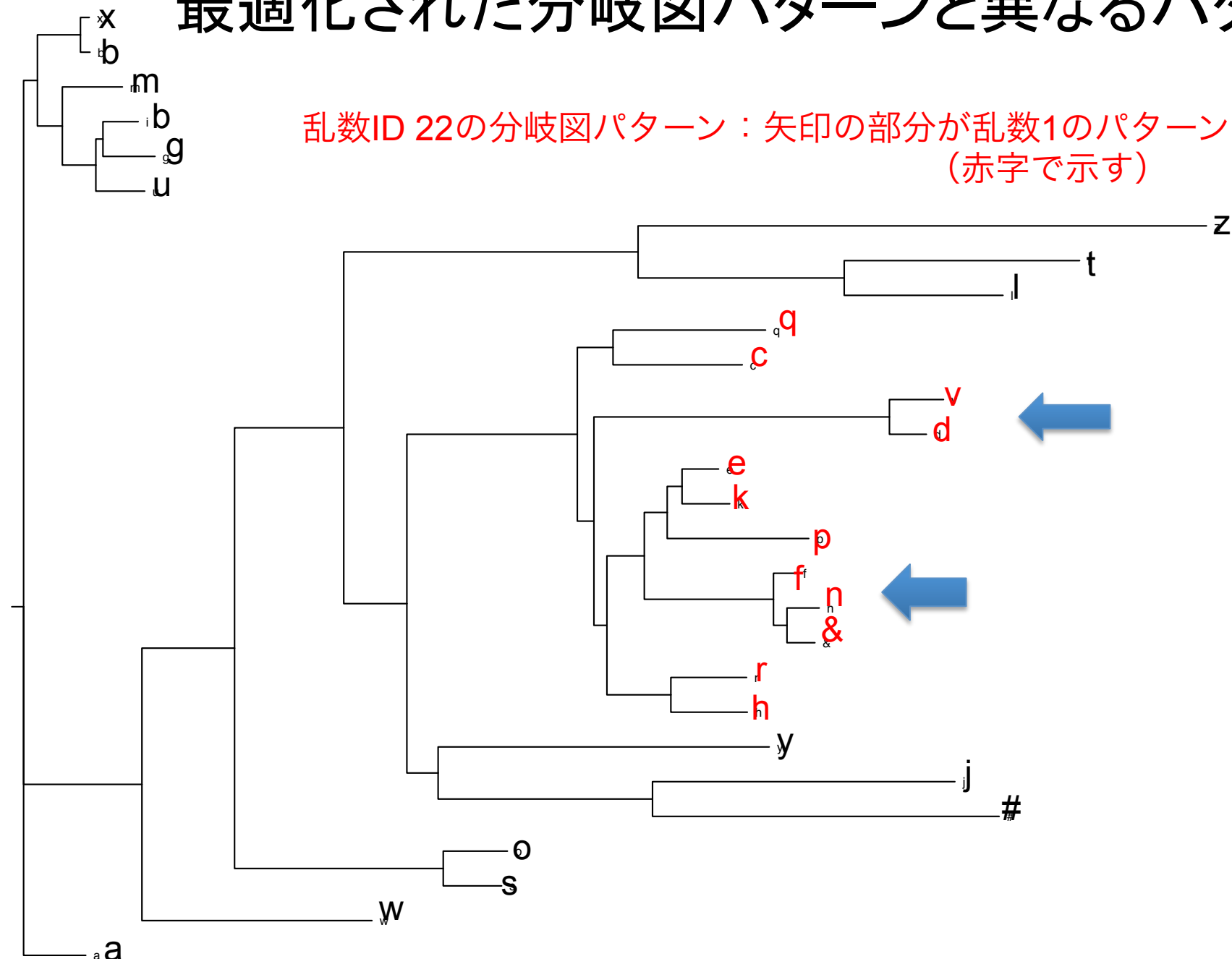
乱数ID 17の分岐図パターン：乱数1のパターンと類似



実施結果:

最適化された分岐図パターンと異なるパターン

乱数ID 22の分岐図パターン：矢印の部分が乱数1のパターンと異なる
(赤字で示す)



(((x,b),(m,((i,g),u))),((((z,(t,l)),(((q,c),((v,d),(((e,k),p),(f,(n,&)))),(r,h))))),(y,(j,#))),(o,s)),w),a)

まとめ:

- 1、**モンテカルロ法による無作為抽出法による分岐図の最適化法を確立した。**
しかし、現状の解析は十分でない。理論上考えられるすべての組合せを計算し、最適化するというのは、まだ困難。さらに工夫が必要
(分散処理によるスケールアウトを含む)。
 - 2、その実施には、大規模なコンピュータリソースを必要とし、
気軽にできるとは言いがたい。
 - 3、今後、ゲノム科学のデータ産生量の増加に伴い、同様の大量の計算機リソース
を必要とする場面が増えてくることが予想される。
- Big Iron** や**クラウド**などを導入したビッグデータ処理環境の需要の高まりが
予想される。

協力者:

東京農工大学(次世代シーケンサー解析)

佐藤暁(2013年度S-PLUS学生研究奨励賞佳作)

有江力、古崎利紀、高橋信弘

茨城大学(分岐図作成の最適化法)

松田朋子(2013年度S-PLUS学生研究奨励賞佳作)

後藤哲雄

アマゾンウェブサービス (クラウド、ビッグデータ)

(AWS in Education Research Grant Award (2013)授与)

ユニバーサル・シェル・プログラミング研究所

(シェルスクリプトによるビッグデータ処理環境共同研究)

その他の事例など:

1、今回も用いた解析サーバのスペックは以下のとおり。

HP ProLiant DL980 G7

CPU: Xeon E7 (80コア160スレッド)

メモリ: 4TB: 32GBDIMMx128枚

HDD: 7.2TB: 900GBx8台

OS: RHEL6.3

2、モンテカルロ法による次世代シーケンサーの配列データの品質評価については、シェルスクリプトベースの分散ファイルシステムによるブルートフォースアプローチ（総当たり計算法）でも解析を実施し、非常に高速で良好な結果を達成した。（すでに8月に成果をMITで発表、論文発表準備中。後日国内でも公開）

3、情報処理学会にて、ビッグデータ活用実務フォーラムを設立。今後ビッグデータに関する情報交換、情報共有を進めていきたい。皆様のご参加を期待する。

Fin



Reminding Waterloo
Bridge (1940)



Powered by
FreeBSD