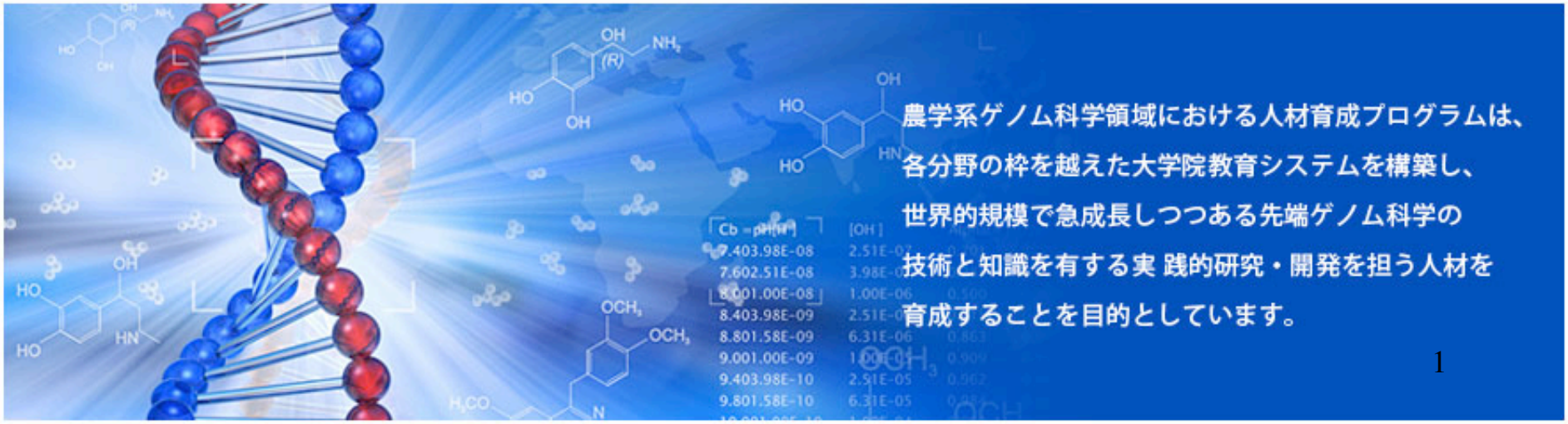
フリーソフトウェアを用いたゲノム科学における ビッグデータ解析

石井一夫
東京農工大学
農学系ゲノム科学人材育成プログラム



自己紹介:プロフィール

石井一夫(東京農工大学特任教授)

専門分野:
ゲノム科学、バイオインフォマティクス、データマイニング、
計算機統計学

経 歴:
徳島大学大学院医学研究科博士課程修了後、
東京大学医科学研究所ヒトゲノム解析センター リサーチアソシエート、
理化学研究所ゲノム科学総合研究センター研究員、
フランス国立遺伝子 多型解析センターCNG研究員、
米国ノースウエスタン大学Feinberg医学部バイオインフォマティクススペシャリストなどを経て現職。

自己紹介：プロフィール

石井一夫（東京農工大学特任教授）

著書など：
著書「図解よくわかる データマイニング」日刊工業新聞社（2004）
翻訳書「ソフトウェアエンジニアリング論文集80‘s～デマルコセレクション」翔泳社（2006）
著書 「統計解析環境Rによるバイオインフォマティクスデータ解析」共立出版（2007）
翻訳書「翻訳バイオエレクトロニクス」 NTS（2008）
翻訳書 「Rによる計算機統計学」オーム社（2011）他。

著書

図解よくわかるデータマイニング（日刊工業新聞社）

ソフトウェアエンジニアリング論文集80(翔泳社)
12章　TeXのエラー、クヌース著を翻訳

著書（R関係）

統計解析環境Rによる
バイオインフォマティクスデータ解析（共立出版）

Rによる計算機統計学（オーム社）

今日の内容

ネット上にすでに要点だけアップしています。

http://www.ospn.jp/press/20130124no32-1-useit-oss.html

OPSN Press メールマガジン
「オープンソース」を使ってみよう (第27回 フリーソフトウェアを用いたゲノム科学におけるビッグデータ処理)

ゲノム科学

次世代シーケンサー、マイクロアレイ、質量分析装置などの装置の普及により、データ産生量が爆発的に増えている。

ギガベース、テラベース級のゲノム解析データを処理して、情報解析を行うことが必要な機会が増えてきた。

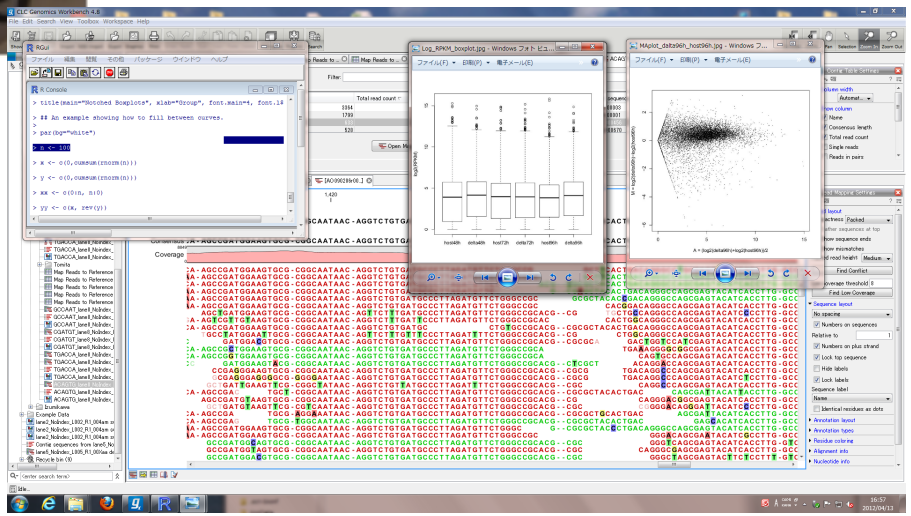
たとえば、次世代シーケンサーのデータだと百塩基程度の配列データとその、クオリティデータで、数千万データとか、数億データとかをテキスト処理するとか、、、

そういうのは普通にある。

次世代シーケンサーを用いたデータ解析

次世代シーケンサー

次世代シーケンサーのデータ解析



ゲノム科学でデータ解析に用いるソフトウェア群

OS　　Linux/UNIX (CentOS 6.3, Scientific Linux 6.3 and FreeBSD 9)
プログラミング言語　　Perl, Python, Ruby, Java, C, C++
データベース　　MySQL, PostgreSQL

ゲノム配列データのアセンブリ Velvet, ABySS, SOAPdenovo,
　　WGS Assembler, MIRA3, Phrap

ゲノム配列データのマッピング　Bowtie, Bowtie2, BWA, SOAP
RNA 発現解析用ソフト Tophat, Cufflinks, Trinity, Oases,
　SOAPdenovo-Trans
ChIP-Seq解析用ソフト　MACS, Quest, SISSRs, SPP

統計解析ソフト　R/Bioconductor, Octave
相同性解析、注釈付けソフト　　BLAST, BLAT
総合DNA解析ソフトウェア　　EMBOSS

生物学データ解析用ライブラリ BioPerl, BioRuby, BioPython, BioJava

ビッグデータ解析用ソフトウェア Hadoop, OpenStack, Eucalyptus

演習用 MacBook Proにプレインストールされた ゲノム解析用フリーソフトウェアの 一覧（一部）

プログラミング言語等
- Perl/BioPerl
- Ruby/BioRuby
- Python/BioPython
- Java/BioJava
- NumPy

データベースソフト
- MySQL

統計解析
- R/Bioconductor
- Octave

ビューア
- IGV
- Tablet

品質チェック
- FastQC

アセンブリソフト
- Velvet
- ABySS
- MIRA3
- Phred/Phrap/Consed
- CAP3

マッピングソフト
- Bowtie
- Bowtie2
- BWA

RNA-Seq解析ソフト
- Oases
- TopHat
- Cufflinks

相同性解析
- BLAST/BLAST+

マルチプルアラインメント
- ClustalW
- ClustalX

モチーフ解析ソフト
- HMMER
- MEME

系統樹解析など
- EMBOSS
- Phylip

ChIP-Seq解析ソフト
- MACS
- SISSRs
- QuEST
- SPP

ビッグデータ解析用ソフト
- Hadoop, Hive

農学系ゲノム科学におけるビッグデータ解析の実施内容

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）

FreeBSD, Linux の操作、インストール、Perlなどをもちいたテキスト処理

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）

Velvet, Oases, Trinity などの操作とデータアセンブリー方法、原理

MySQL, PostgreSQL を用いたデータベースの構築と、クエリ、集計

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）

発現定量データの取得と統計解析、パラメトリック検定、ノンパラメトリック検定、多変量解析、機会学習、クラスター解析、グラフ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）

遺伝統計解析、統計モデリング（一般化線形モデル、一般化加法モデルなど）、モンテカルロシミュレーション、マルコフ連鎖モンテカルロ法、遺伝学的系統樹解析

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）

Perl, Python, Ruby, C，C++，Javaを用いた新規アルゴリズムの実装。

1. ゲノム科学で用いられるフリーソフトウェア

次世代シーケンサーというDNA塩基配列情報を
大量に産生する機器が実用化されて数年が経過し、
ゲノム科学におけるデータ産生量や、その取り扱う
データ量が飛躍的に
増えています。

これらのデータ処理にはUNIX/Linuxを中心とする
フリーソフトウェアは欠かせません。

（1）汎用のフリーソフトウェア

特に、次世代シーケンサーのデータは、
例えばイルミナ社のデータですと1ファイルあたりに
数千万断片から数億断片のDNA塩基配列データと
そのクウォリティデータが産生されます。

それを、
(1) catやgrep、sed、awkなどのシェルのコマンドや、
(2) Perl、Python、Rubyなどのスクリプト言語、
(3) R、Octaveなどの統計解析言語を組み合わせて
処理します。必要に応じて、
(4) MySQL、PostgreSQLなどのデータベースも使用
します。むしろ、このようなスケールのデータ解析では、
データベースは必須です。

次世代シーケンサーからのデータ例
Fastqファイル

@HWUSI-EAS1748:79:66020AAXX:4:1:8488:1047 1:N:0:ATGTCA
AAGCATTCTAAGGCGAAGCCACCCATTCTTTCCTGCATATATACTTACAAACACATAGCCCCCATCT
+
HHHHHHHHHHHHHHHHHHHHHHHHHAHGHHHHHDFHHHHHHGHHHHHHHHHHHHHHHHHHF
@HWUSI-EAS1748:79:66020AAXX:4:1:14340:1047 1:N:0:ATGTCA
AAGCTTTCTGGTGATCGACGCGCATGGCCATGAGGAGGACTCGTCGCCGGATCATCCTCCCGTTTC
+
IIGIIIIIIIIIIIIIIIIHIHIIIIGHEBI@GGGFGHIHGHIIFIIGBIFEIFFIIHFGHHHHGG2C@CEEDA=BD@D
@HWUSI-EAS1748:79:66020AAXX:4:1:17830:1047 1:N:0:ATGTCA
GGAAATTTAAGCGACCACGAAGAGTATGACGCTGGTGAAGATTGGTCCGTGGGGCGGAAATGGA
+
IIIIFIIIIIIIIIIIIIIGIFIIIIIIIIIEIIIIIIIIGIIIGIHIHHIICGHIIHEIGFADEHFEE@GGADEBD>EAC:CC@9
@HWUSI-EAS1748:79:66020AAXX:4:1:2618:1047 1:N:0:ATGTCA
ATTAAGAAGAGAAGGCACTTGTCAGATGGTTCGAAGCATATGCTTACTGAAATGGAGAGAGCAGA
+
FIIIIGBGGGIEIIGIIHIIIIIFIBGGGEGGG@GIIIEGEGIIIDIGIEGDGDGAF@FAGGGGDGG>EGGB>DD
@HWUSI-EAS1748:79:66020AAXX:4:1:17486:1047 1:N:0:ATGTCA
AAAAAAATAGAAATCTCTCACTATGAAGTATGAACTCAACATGGACTATCATACGACCATCATGCC
+
DHHGHHHHHGBGHHHHHGHGHHHHHHDHHHHHHHGHHHHHHHHHHGHHHHHHHHBBHGHGH
@HWUSI-EAS1748:79:66020AAXX:4:1:5559:1047 1:N:0:ATGTCA
GACCCAAGGGCAGCAGCAGGGCTACTCTCAGCAGACTGGATATGATCAGTAGGGCTATGGAACTT

次世代シーケンサーからのデータ例
Fastqファイル

以下のようなデータが数千万行から一億行あるいはそれ以上

@HWUSI-EAS1748:79:66020AAXX:4:1:8488:1047 1:N:0:ATGTCA
AAGCATTCTAAGGCGAAGCCACCCATTCTTTCCTGCATATATACTTACAAACACATAGCCCCCATC
+
HHHHHHHHHHHHHHHHHHHHHHHHHHAHGHHHHHDFHHHHHHGHHHHHHHHHHHHHHHHHH
@HWUSI-EAS1748:79:66020AAXX:4:1:14340:1047 1:N:0:ATGTCA
AAGCTTTCTGGTGATCGACGCGCATGGCCATGAGGAGGACTCGTCGCCGGATCATCCTCCCGTTT
+
IIGIIIIIIIIIIIIIIIIIHIHIIIIGHEBI@GGGFGHIHGHIIFIIGBIFEIFFIIHFGHHHHGG2C@CEEDA=BD@I
@HWUSI-EAS1748:79:66020AAXX:4:1:17830:1047 1:N:0:ATGTCA
GGAAATTTAAGCGACCACGAAGAGTATGACGCTGGTGAAGATTGGTCCGTGGGGCGGAAATGGA
+
IIIIFIIIIIIIIIIIIIIGIFIIIIIIIIIIEIIIIIIIIIGIIIGIHIHHIICGHIIHEIGFADEHFEE@GGADEBD>EAC:CC@9
@HWUSI-EAS1748:79:66020AAXX:4:1:2618:1047 1:N:0:ATGTCA
ATTAAGAAGAGAAGGCACTTGTCAGATGGTTCGAAGCATATGCTTACTGAAATGGAGAGAGCAG
+
FIIIIGBGGGIEIIGIIHIIIIIFIBGGGEGGG@GIIIEGEGIIIDIGIEGDGDGAF@FAGGGGDGG>EGGB>DI
@HWUSI-EAS1748:79:66020AAXX:4:1:17486:1047 1:N:0:ATGTCA
AAAAAAATAGAAATCTCTCACTATGAAGTATGAACTCAACATGGACTATCATACGACCATCATGC

次世代シーケンサーからのデータ例
Fastqファイル

1行目 配列名

2行目 DNAの塩基配列

```
@HWUSI-EAS1748:79:66020AAXX:4:1:8488:1047 1:N:0:ATGTCA
AAGCATTCTAAGGCGAAGCCACCCATTCTTTCCTGCATATATACTTACAAACACATAGCCCCCATCT
+
HHHHHHHHHHHHHHHHHHHHHHHHHHAHGHHHHHDFHHHHHHGHHHHHHHHHHHHHHHHHHF
```

4行目 DNAの塩基のクオリティ

テキスト処理のための Perlスクリプトの例

```
#!/usr/bin/perl
# Name: fastq-SeqExtract.pl
# usage:
# perl fastq-SeqExtract.pl input_filename start end output_filename
# or
# chmod +x fastq-SeqExtract.pl
# ./fastq-SeqExtract.pl input_filename start end outoutfilename
# by Kazuo Ishii, Ph.D., NOV 29. 2012

$input = $ARGV[0];
$start = $ARGV[1];
$end = $ARGV[2];
$output = $ARGV[3];

open ( FILEHANDLE ,  "< $input") ;
open ( FILEHANDLE2 , "> $output" ) ;

@array = <FILEHANDLE> ;
chomp(@array);
chomp($start);
chomp($end);

$start1 = $start-1;
$end1 = $end - $start;


for (my $i = 0; $i <= $#array; $i += 4){
   $array[$i+1] = substr($array[$i+1],$start1,$end1);
   $array[$i+3] = substr($array[$i+3],$start1,$end1);
   print FILEHANDLE2 "$array[$i]_$input-$start-$end¥n$array[$i+1]¥n$array[$i+2]¥n$array[$i+3]¥n";
          }

close(FILEHANDLE);
close(FILEHANDLE2);
```

生物学的なデータ解析専用のソフトウェア

もちろん、生物学的な情報解析専用のソフトウェアも多数開発されています。例えば、

(1) 配列データのアセンブリには、Velvet、Oases、Trinity、
(2) 配列データの既知の配列へのマッピングには、BWA、Bowtieなどが、
(3) 得られた配列データと既知の配列との相同性検索には、BLASTなどが用いられます。
Perl、Python、Ruby、Javaなどには、
(4) 生物学的解析に特化した関数などを集めたライブラリが整備されており、それぞれBioPerl、BioPython、BioRuby、BioJavaと呼ばれています。また、
Rには、(5) Bioconductorと呼ばれる生物学的解析用のパッケージ群が存在します。

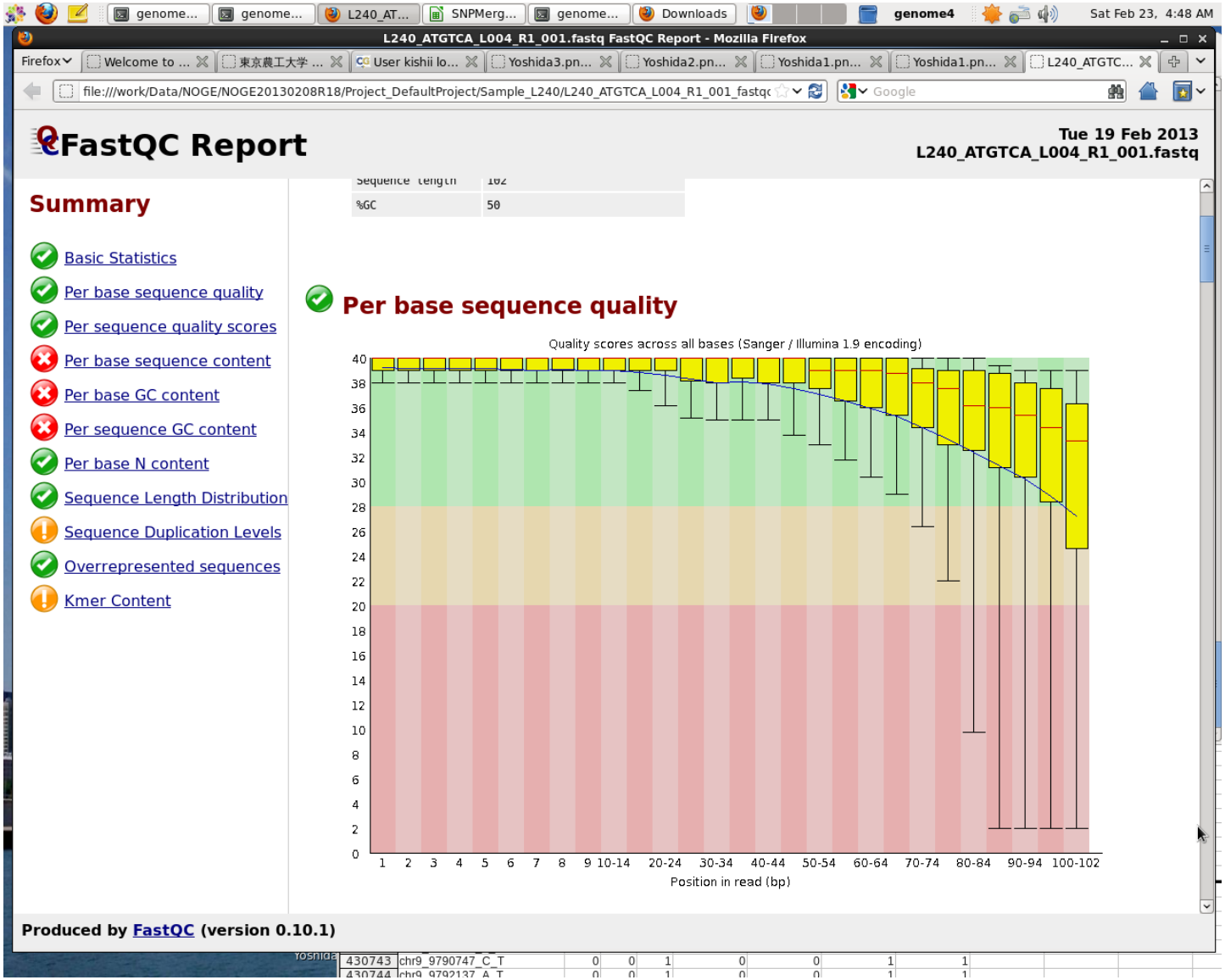
次世代シーケンサーの解析ワークフロー　RNA-Seq の場合

1. ローデータ
 bclファイル,fastqファイル
2. クウォリティチェック
  FastQC -> フィルタリング、トリミング
        Cutadapt, Perl スクリプト
3. アセンブリ Velvet, Oases, Trinity, SOAP-denovo….
4. マッピング　BWA, Bowtie, (Maq), TopHat, Cufflinks など

次世代シーケンサーの解析ワークフロー　RNA-Seq の場合

1. ローデータ
 bclファイル,fastqファイル
2. クウォリティチェック
  FastQC -> フィルタリング、トリミング
        Cutadapt, Perl スクリプト
3. アセンブリ Velvet, Oases, Trinity, SOAP-denovo….
4. マッピング　BWA, Bowtie, (Maq), TopHat, Cufflinks など

FASTQCによるクオリティチェック



次世代シーケンサーの解析ワークフロー　RNA-Seq の場合

- Viewer で閲覧

Samtools などを用いて、インデックスを作成する。

IGV, Tablet などで閲覧。

次世代シーケンサーの解析 ワークフロー RNA-Seq の場合

5. カウントデータの採取

（1）SamファイルをBedファイルに変換し、R/Bioconductorパッケージでカウントデータ、RPKMを算出、MA-plot を書く。

（2）Samファイルをawk, grep で整形し、sort, uniq などでカウントデータ、RPKMを算出。

（3）SamファイルをMySQLデータベースに格納し、

MySQLを用いて、カウントデータを算出、RPKMを算出。

次世代シーケンサーの解析ワークフロー　RNA-Seq の場合

- データマイニング

いわゆる統計解析とか、クラスター解析とか、

PCAとか、マイクロアレイでやられていた方法に持ち込む。

•パスウェイ解析

GOとか、KEGG とか、

GESAとか。。

次世代シーケンサー解析
ワークフロー　RNA-Seqの場合

6. アノテーション

BLAST検索。→結果を、

(1)シェルスクリプトや Perl などで、抽出。

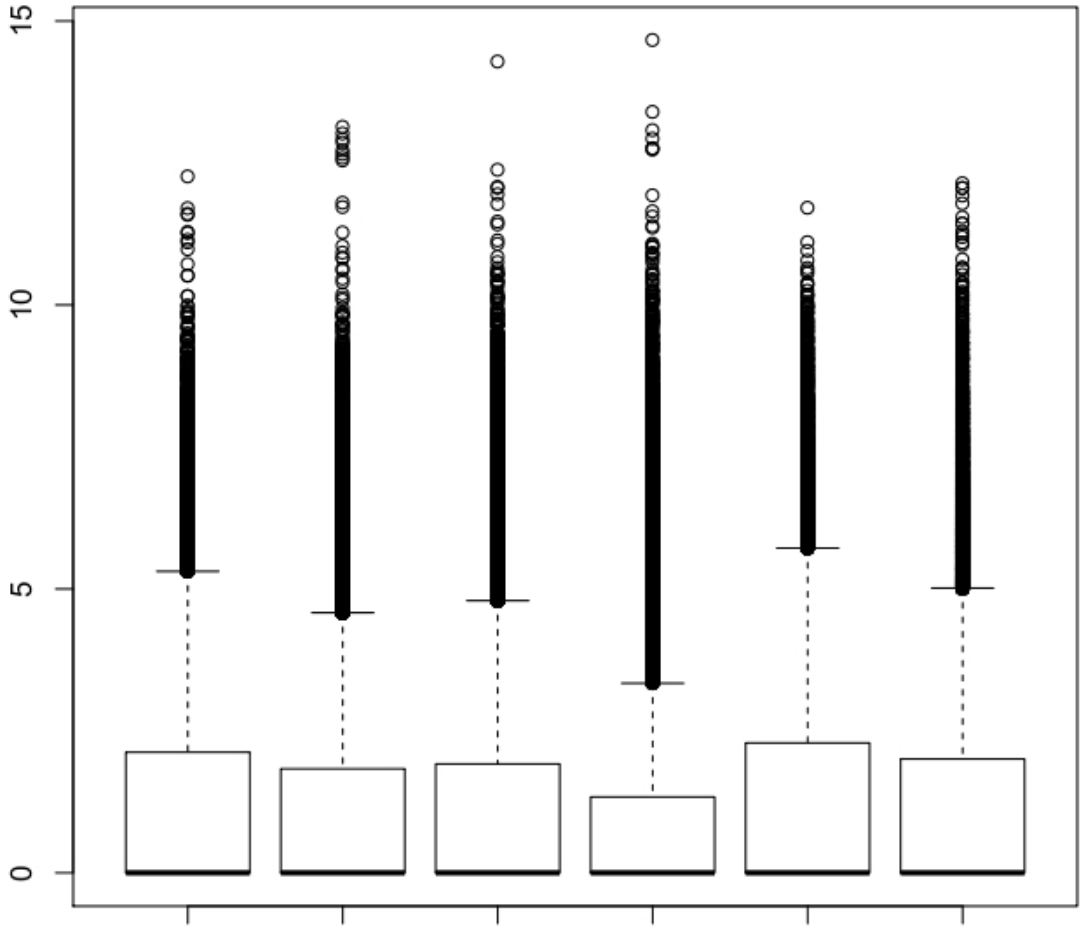
(2)データベースに収納し、SQLで抽出。

して、整理する。

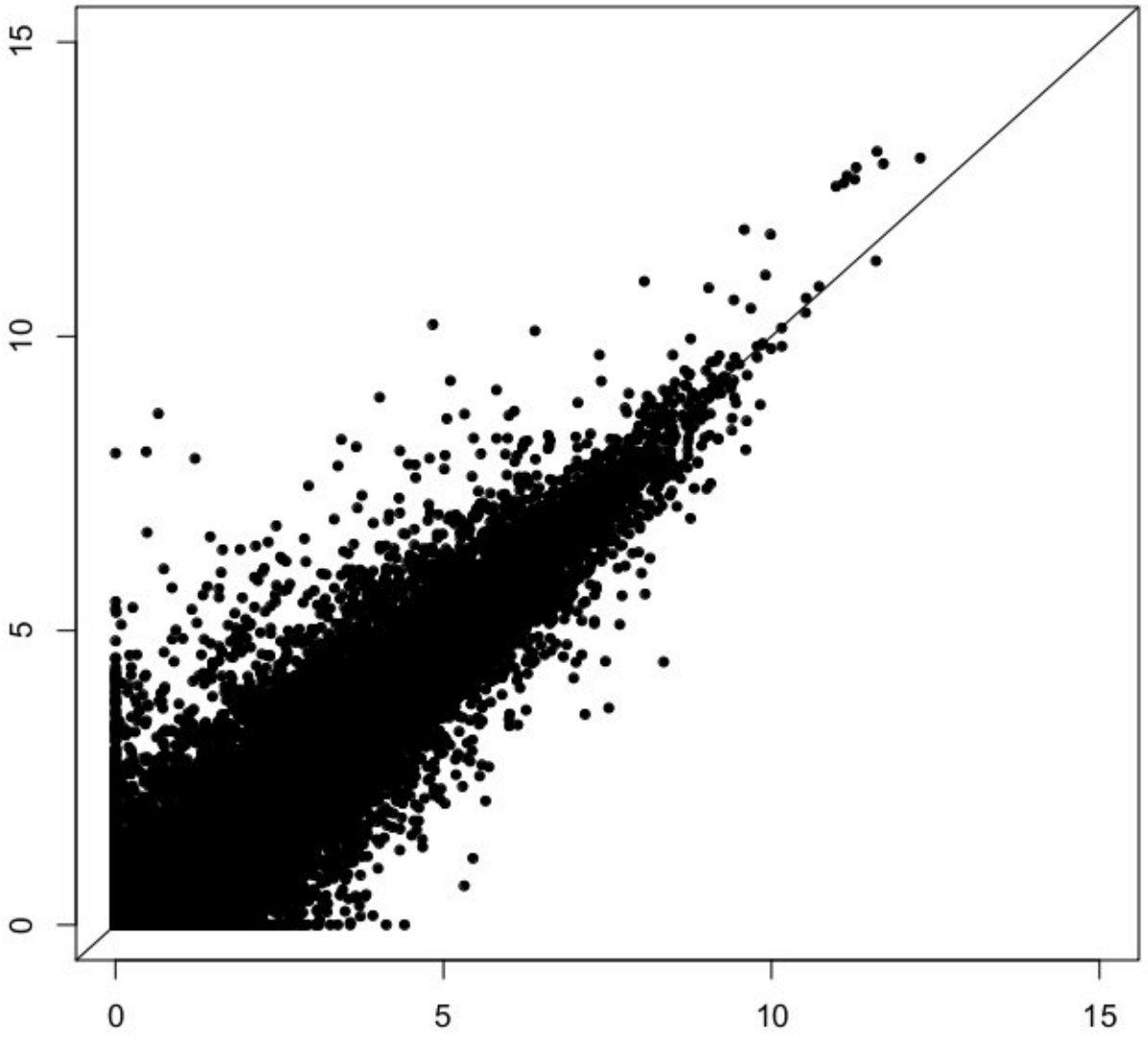
7. グラフ作成

8.統計解析、データマイニング

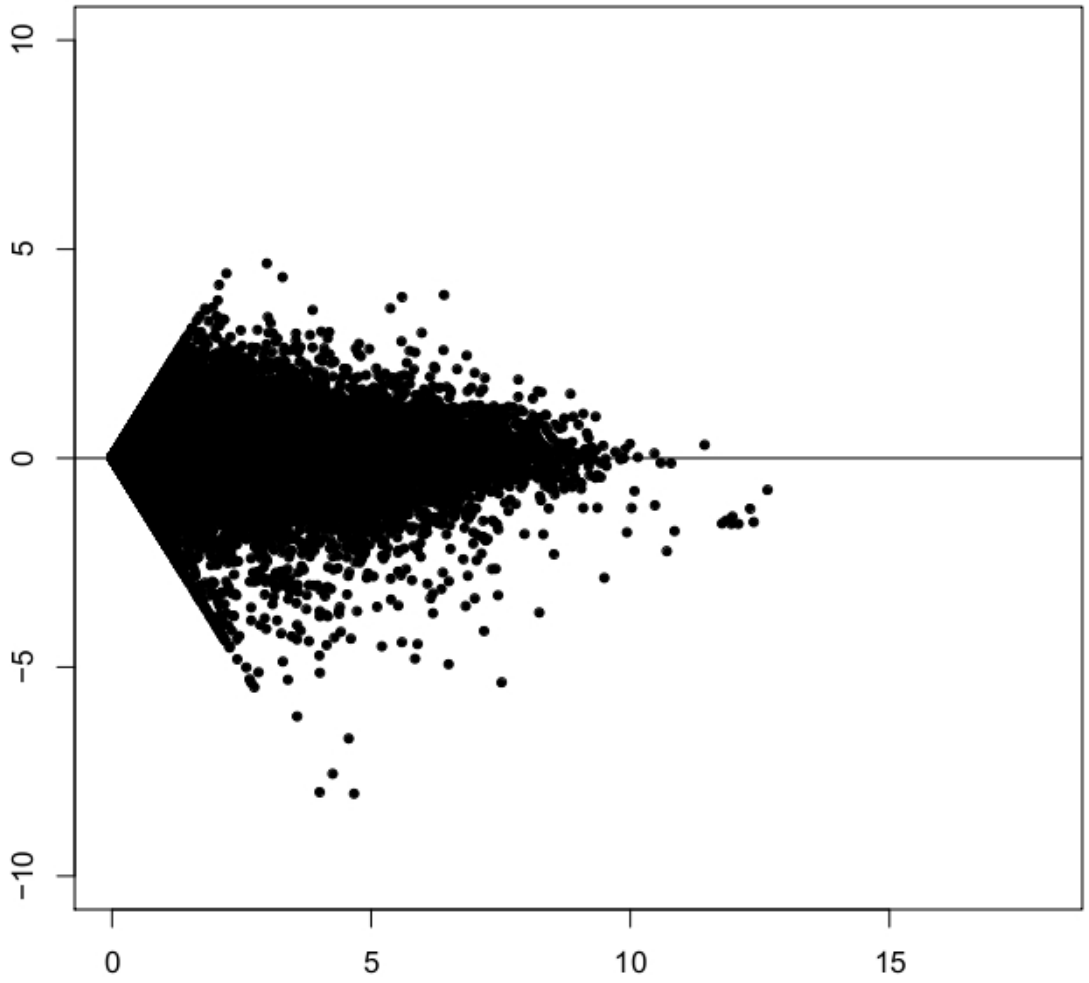
Rでグラフ作成



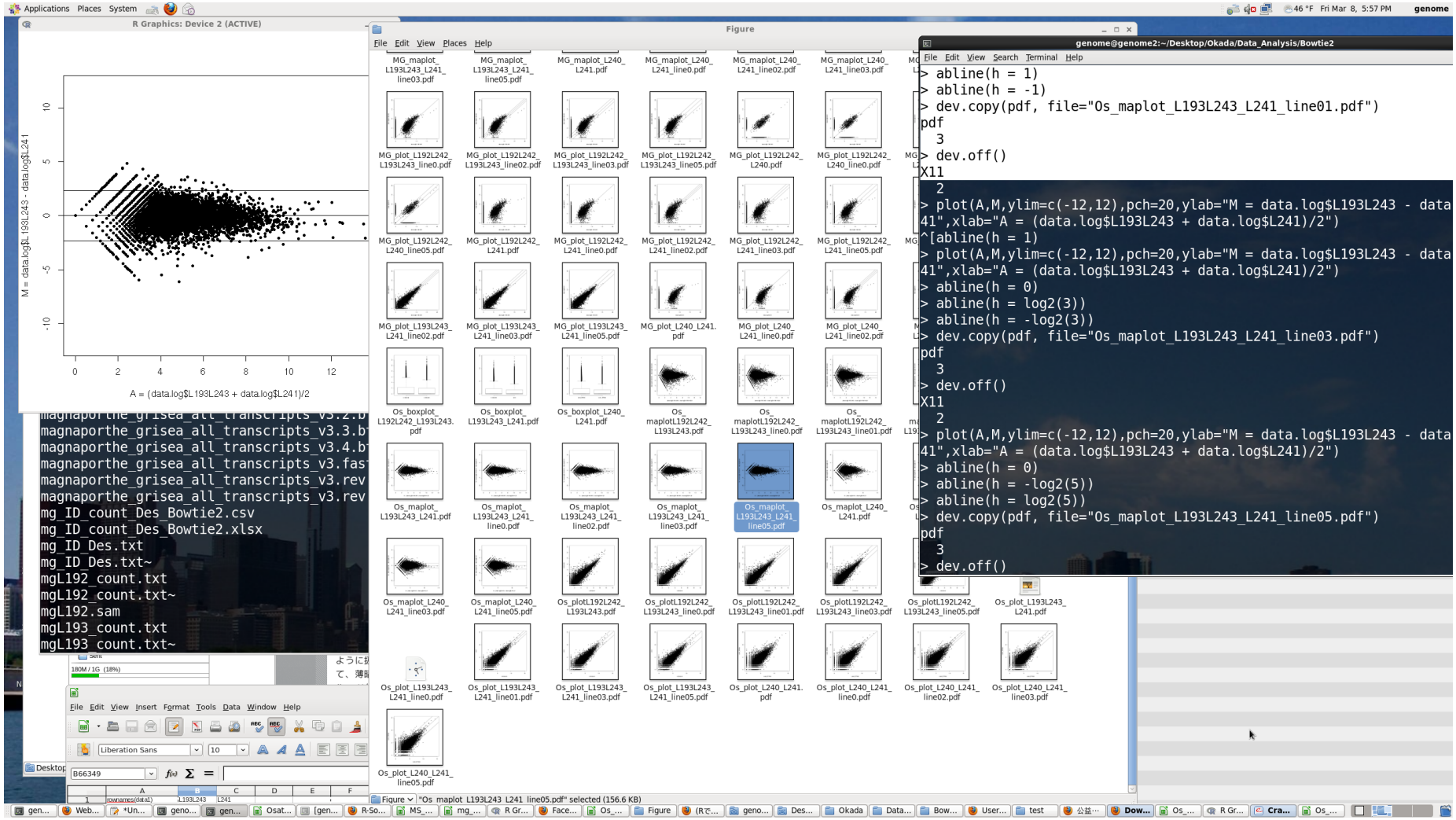
Rでグラフ作成



Rでグラフ作成



実際の解析風景


R Graphics: Device 2 (ACTIVE)
M = data.log$L193L243 - data.log$L241
A = (data.log$L193L243 + data.log$L241)/2
Figure
genome@genome2:~/Desktop/Okada/Data_Analysis/Bowtie2
> abline(h = 1)
> abline(h = -1)
> dev.copy(pdf, file="Os_maplot_L193L243_L241_line01.pdf")
> dev.off()
> plot(A,M,ylim=c(-12,12),pch=20,ylab="M = data.log$L193L243 - data
> abline(h = 0)
> abline(h = log2(3))
> abline(h = -log2(3))
> dev.copy(pdf, file="Os_maplot_L193L243_L241_line03.pdf")
> abline(h = -log2(5))
> abline(h = log2(5))
> dev.copy(pdf, file="Os_maplot_L193L243_L241_line05.pdf")
magnaporthe_grisea_all_transcripts_v3.2.b
magnaporthe_grisea_all_transcripts_v3.3.b
magnaporthe_grisea_all_transcripts_v3.4.b
magnaporthe_grisea_all_transcripts_v3.fas
magnaporthe_grisea_all_transcripts_v3.rev
magnaporthe_grisea_all_transcripts_v3.rev
mg_ID_count_Des_Bowtie2.csv
mg_ID_count_Des_Bowtie2.xlsx
mg_ID_Des.txt
mg_ID_Des.txt~
mgL192_count.txt
mgL192_count.txt~
mgL192.sam
mgL193_count.txt
mgL193_count.txt~

次世代シーケンサーの解析ワークフロー　ChIP-Seq の場合

Read のmapping

1. ローデータ
   bclファイル,fastqファイル
2. クウォリティチェック
   FastQC -> フィルタリング、トリミング
   Cutadapt, Perl スクリプト
3. アセンブリ Velvet, Oases, Trinity, SOAP-denovo….
4. マッピング BWA, Bowtie, (Maq), TopHat, Cufflinks など

基本的にRNA-Seqとよく似ています。

次世代シーケンサーの解析ワークフロー　ChIP-Seq の場合

- Viewer で閲覧

Samtools などを用いて、インデックスを作成する。

IGV, Tablet などで閲覧。

次世代シーケンサー解析 ワークフロー ChIP-Seq

5. ピークコーリング Peak calling

SISSRs, MACS, PeakSeq, QuEST, FindPeaks,SPP, CisGenomeなどでピークの位置を特定。

6. ピーク位置から、ピークの配列を抽出し、クラスタリング。

7.モチーフ解析

Motif enrichment analysis in sequences under peaks

次世代シーケンサー解析 ワークフロー ChIP-Seq

6. アノテーション Peak annotation/Filtering
ChIPPeakAnno などで検索。→結果を、
(1)シェルスクリプトや Perl などで、抽出。
(2)データベースに収納し、SQLで抽出。
して、整理する。
7. グラフ作成
8.統計解析、データマイニング
Differential peak analysis

次世代シーケンサー解析 ワークフローまとめ ChIP-Seq

1. Read mapping
2. Peak calling
3. Peak annotation/Filtering
4. Differential peak analysis
5. Motif enrichment analysis in sequences under peaks

ピークコーリング

- ピークコーリングソフト

CisGenome, ERANGE, FindPeaks, F-Seq, GLITR, MACS, PeakSeq, QuEST, SICER, SiSSRs, spp, Useq etc

ChIP-Seq に用いられるRパッケージ

- GenomicRanges ：遺伝子の特定範囲の取り扱い
- Rsamtools ：BAM の処理
- Rtracklayer ：ゲノムブラウザからのアノテーション情報取り込み
- DESeq ：RNA-Seq
- edgeR ：RNA-Seq
- Chipseq ：ChIP-Seq 解析ソフト
- ChIPpeakAnno : ゲノム情報によるピークのアノテーション

多くの解析は、Rを使わなくてもSAMToolsなどで代用可能

ピークコーリングなどに用いられるRパッケージ

- BayesPeak : 隠れマルコフモデルとベイズ統計
- PICS : ChIP-Seqの確率推論
- DiffBind Link : ChIP-Seq ピークデータの結合差解析Overlap 計算、Boxplot, PCA biplot, heatmap による可視化、edgeR, DESeq をつかった binding affinity 解析
- MOSAiCS :モデルから期待される値に fitting
- iSeq Link :隠れイジングモデルによる結合部位の同定
- ChIPseqR :タンパク結合部位とヌクレオソーム位の推定
- CSAR: ポアソン分布による検定
- ChIP-Seq ：SPPを用いるChIP-Seq 解析パイプライン
- SPP:ピークコーリング

統計解析

- マイクロアレイが比率データであるのに対し、次世代のデータは、カウントデータである。
- このため負の二項分布（又は、ポアソン分布）を示す。
- ゼロ値をもつ遺伝子が極端に多いことも考慮する必要あり。

BioPerl, BioRuby などによる解析

- BioPerl, BioPython, BioRuby, BioJava とは
- Perl, Python, Ruby, Java などで、生物学的情報処理に用いられる専用ライブラリ
- 検索などを自動化

使用例(BioRuby)

- Genbank GeneID(Acc) リストから、アノテーション情報を自動取得
- #!/usr/bin/env bioruby
  # bioruby script.rb > out.txt

```
File.open("ListGenbank.txt").each do |item|
Bio::NCBI.default_email = "kishii@cc.tuat.ac.jp
  entry = getent("genbank:#{item}")
  gb = flatparse(entry)
  item1 = item.chomp
  print item1, "\t", gb.definition, "\t", gb.organism, "\n"
end
```

使用例(BioPerl)

- Genbank GeneID(Acc) リストから、アノテーション情報を自動取得
- #!/usr/bin/perl -w
  use Bio::DB::GenBank;
  use Bio::Species;

  my $file = $ARGV[0];
  open(my $fh, "< $file") or die "Cannot open $file: $!";

  while (my $line = readline $fh) {
    chomp $line;
    $db_obj = Bio::DB::GenBank->new;
    $seq_obj = $db_obj->get_Seq_by_acc($line);
    print $seq_obj->accession, "\t", $seq_obj->desc, "\n";

  }

今後の課題

（1）ビッグデータ解析のための教科書など
教材の整備。
（2）ビッグデータ解析のためのデータベースなどの整備。
各種アノテーションデータベースの整備
（3）クラウドによるビッグデータ解析環境の整備。
Hadoop, OpenStackなど