

# S-PLUS/R による 次世代シーケンサーの データ解析

東京農工大学  
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

# Agenda

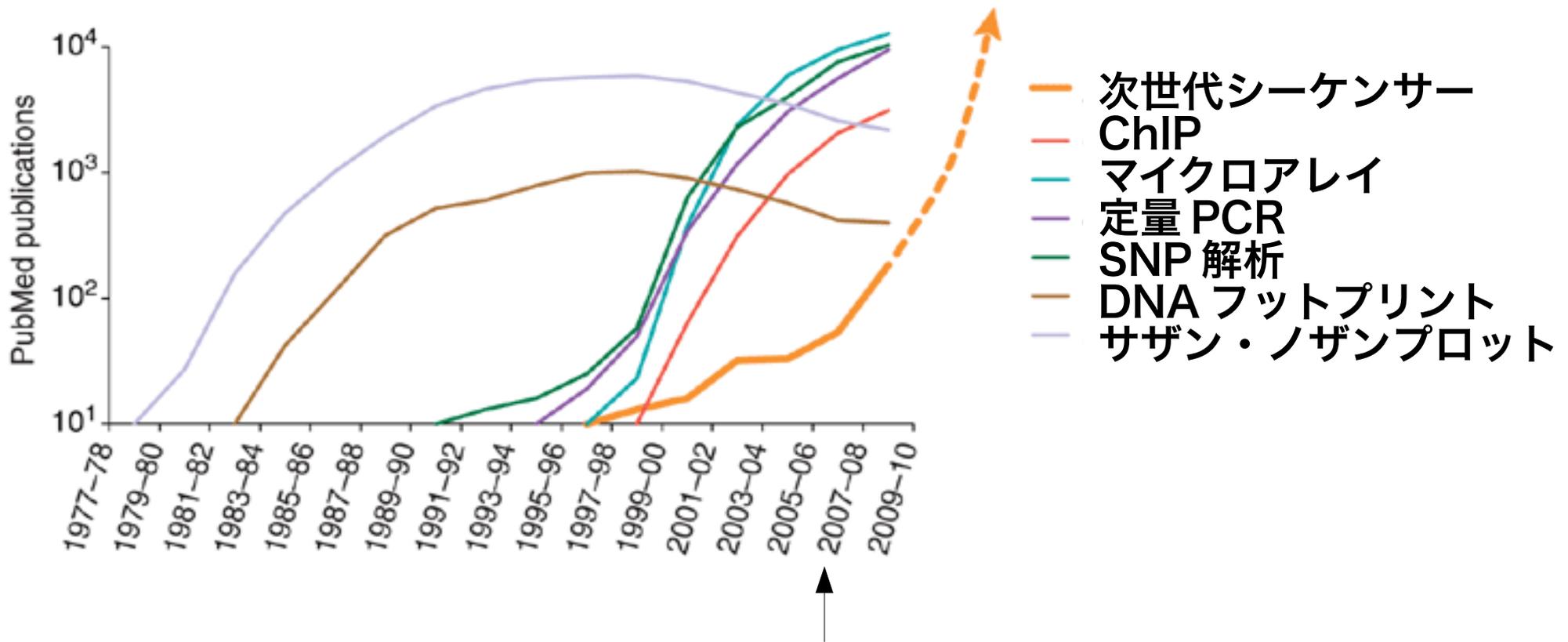
- 次世代シーケンサーとは
- 次世代シーケンサーの応用
- 農学分野におけるゲノム科学の応用とその特徴
- 次世代シーケンサーによるデータ解析
- S-PLUS/Rによるデータ解析
- （デモ）今日は無理かも、、。

**次世代シーケンサーとは**

# 次世代シーケンサーとは

- 従来のサンガー法によらない大規模並列塩基配列決定法による自動解析装置。
- DNA ポリメラーゼまたは DNA リガーゼを用いて逐次的 DNA 合成法を行い、蛍光・発光などの光検出により、超並列的に塩基配列を決定。最近では、イオン電荷を半導体で検出する安価なものも登場。
- 1～25 Gb/日
- 価格 1 千万円～1 億円
- 25～400 塩基程度（それ以上のものも出現）
- 数百万～数十億リード / 回
- ヒトゲノムを 1 週間程度で解析可能と言われる。
- **ゲノム解析・ゲノムデータ解析を研究室レベルで可能にした。**

# 次世代シーケンサーの論文数の推移



次世代シーケンサーの実用化、普及し始めた  
2008年後度から急速に論文数が増加

# 次世代シーケンサー

## 例) HiSeq 2000 (Illumina 社)

- 従来のサンガー法によらない大規模並列塩基配列決定法 (SBS 法) による自動解析装置。
- スループット 540-600Gb
- リード数 50 億～60 億
- リード長 100 塩基 X 2
- 用途：RNA 発現量の網羅的解析 (RNA-Seq)、ゲノム修飾の網羅的解析 (ChIP-Seq) に向いている。多型解析、ゲノム塩基配列解析にも用いられる。



解析原理：<http://www.youtube.com/watch?v=77r5p8lBwJk>

# 次世代シーケンサー

## 例) 454 GS FLX (ロシュ社)

- 従来のサンガー法によらない大規模並列塩基配列決定法（パイロシーケンス法）による自動解析装置。
- スループット 400-600Mb
- リード数 100万リード
- リード長 500塩基
- 用途：新規ゲノム解析、メタゲノム。



解析原理：<http://www.youtube.com/watch?v=bFNjxKHP8Jc>

# 次世代（半導体）シーケンサー例 Ion BGM（ライフテクノロジーズ社）

- 光学的検出器によらない次世代シーケンサー、パイロシーケンス法により生じたイオン電荷を検出する。
- スループット 400-600Mb
- リード数 100万リード
- リード長 500塩基
- 用途：新規ゲノム解析、メタゲノム、DNA多型解析、RNA発現解析。

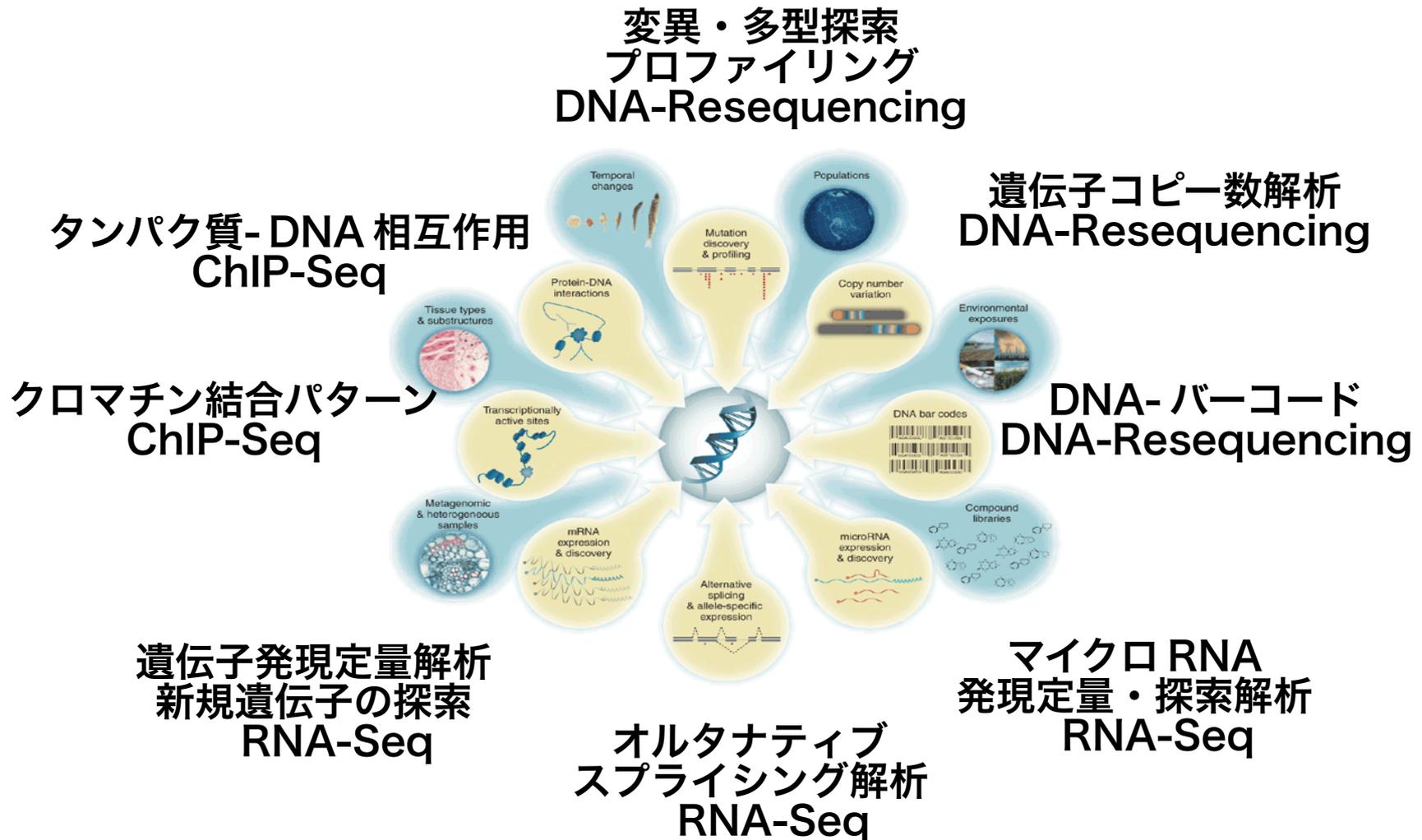


解析原理：<http://www.youtube.com/watch?v=yVf2295JqUg>

# 次世代シーケンサーの応用

# 次世代シーケンサーの多様な用途

新規塩基配列解析以外に考えられる用途



# 3つの主要なアプリケーション

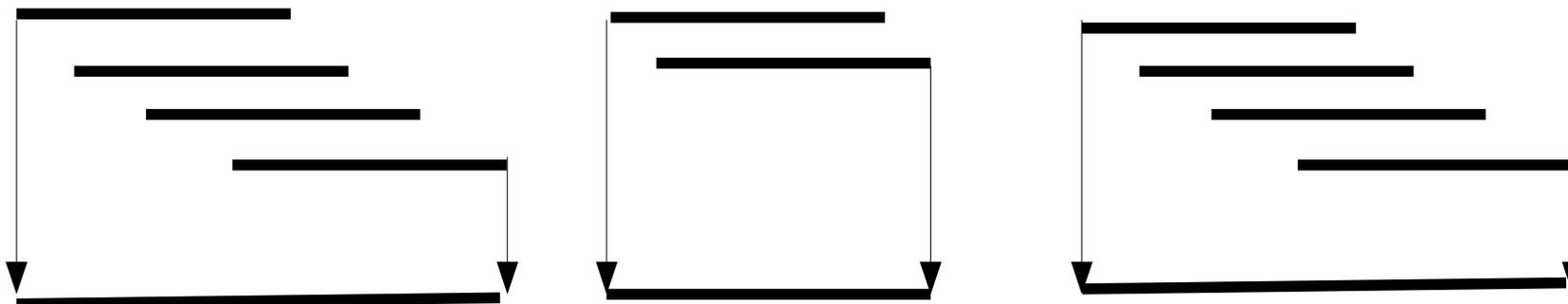
- DNA-Sequencing  
*de novo* Sequencing  
Resequencing (多型解析、遺伝子コピー数解析)  
メタゲノム
- RNA-Sequencing  
*de novo* RNA-Seq - 発現定量・新規遺伝子同定  
RNA-Seq - 発現定量・新規遺伝子同定  
メタゲノム
- タンパク質-DNA 相互作用  
ChIP-Seq - クロマチン、RNA ポリメラーゼ、転写因子

# DNA Sequencing

- マッピング 参照配列にアラインメント  
リシーケンシング・多型解析

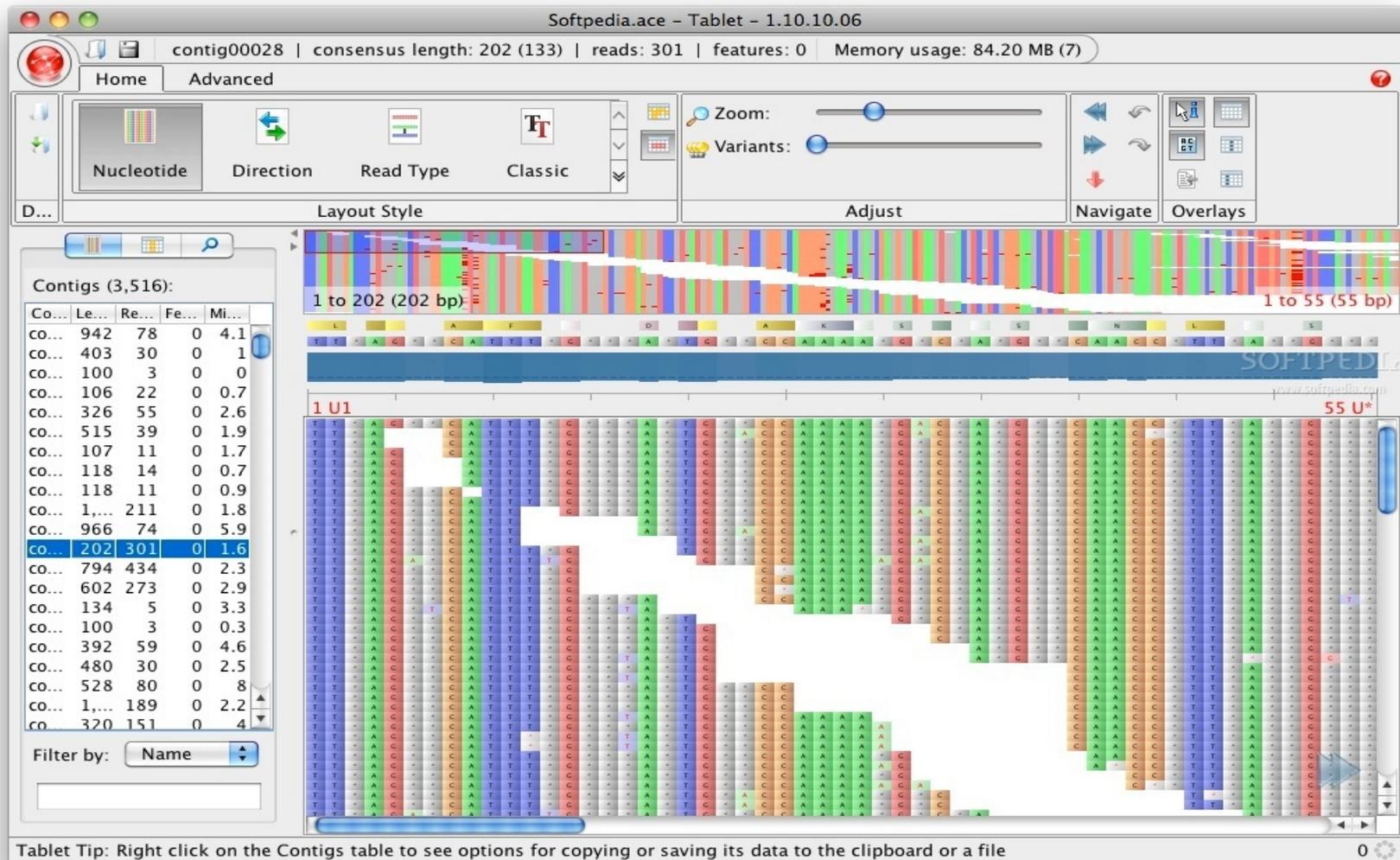


- アセンブリ 塩基配列を重ねあわせコンセンサス配列を得る 新規配列解析 (de novo)



# DNA Sequencing

## Tablet をもちいたマッピング結果の表示例



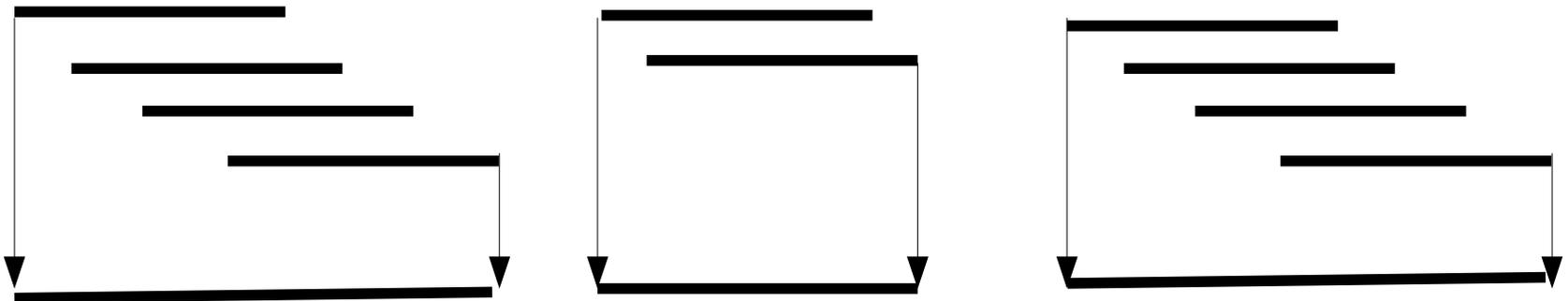
(<http://mac.softpedia.com/get/Math-Scientific/>)

# RNA Sequencing

- RNA-Seq 発現定量解析、新規遺伝子探索、  
低分子量 RNA 同定・定量



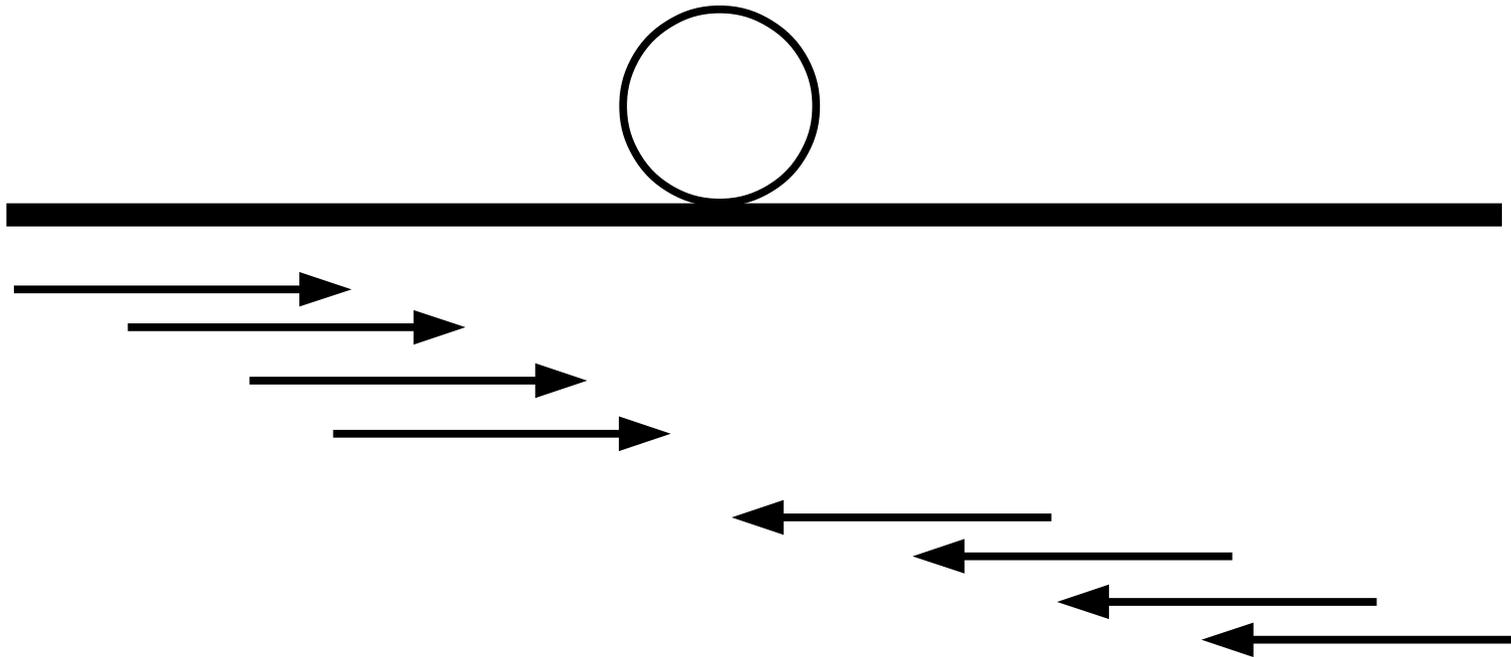
- de novo RNA-Seq 新規遺伝子探索、  
発現定量解析



# ChIP-Seq

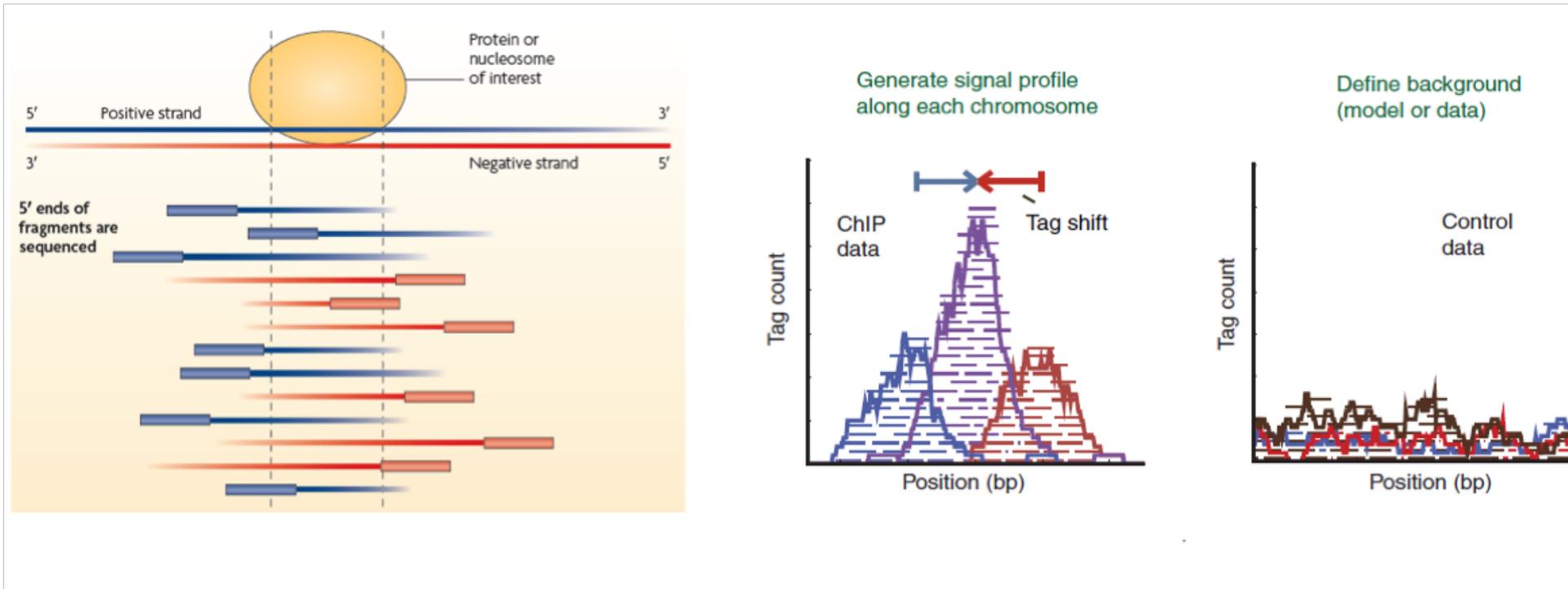
- タンパク質に結合した DNA を免疫沈降で捕捉し、配列解析を行い結合部位を同定。

DNA 結合タンパク質  
クロマチンなど



# ChIP-Seq

- タンパク質に結合した DNA を免疫沈降で捕捉し、配列解析を行い結合部位を同定。



Park, P.  
*Nature Reviews Genetics*, **2009**, *10*, 669-680

Pepke, S.; Wold, B. & Mortazavi,  
*Nature methods*, **2009**, *6*, S22-S32

# • 農学分野におけるゲノム科学の応用とその特徴

# 東京農工大学 「農学系ゲノム科学人材育成プログラム」設置の 経緯と農学分野のゲノム解析

[Japanese](#) [English](#)



東京農工大学 文部科学省「農学系ゲノム科学人材育成プログラム」で使用する主要機器

次世代シーケンサーイルミナGAIIx（左）、ABI5800 MALDI-TOF-TOF質量分析装置（中央）、  
サーモフィッシャーLC-LTQ-Orbitrap質量分析装置（右）

[ホーム](#)

[沿革](#)

[概要](#)

[運営組織・組織内容](#)

[プログラム採択者](#)

[お知らせ・新規募集](#)

[セミナー・講習会](#)

お知らせ

2012年1月18日 **新規**

[第9回農学系ゲノム科学人材育成セミナー](#)

「次世代シーケンサーの最前線：ロシュ・ダイアグノスティクス社」

2011年12月7日 **新規**

[第8回農学系ゲノム科学人材育成セミナー](#)

「バイオインフォマティクス入門～BLASTから次世代シーケンサー解析まで～：（株）  
アメリカエフ」

# 経緯

- 全国遺伝子施設会議による（2008年11月）全国遺伝子実験施設におけるゲノム科学関連機器の導入状況、必要性に関する調査で、具体的研究テーマを有するにも関わらず、全国的にゲノム科学設備の導入は農学系ではほとんどなされていないこと、その設備を運営していく人材がいなかったことが判明。
- これらの調査結果を基に、全国19大学の農学系博士課程を対象としたゲノム科学拠点構想を策定し、ゲノム科学設備導入と博士課程学生のゲノム科学分野での研究推進と人材育成を骨子とした案を立案。
- 2009年度に文部科学省補正予算により次世代ゲノムアナライザー、極微量液体クマトグラフィ-質量分析システム及びマトリックス支援レーザー脱離イオン化飛行型タンデム質量分析システム、リアルタイムレーザー共焦点顕微鏡など、ゲノム科学解析設備を導入。学術研究支援センターで設備の運用及び稼働体制を整えた。
- 2011年度から文部科学省特別経費で採択され、5年計画の本プログラムを全体総括・松永是学長、プログラム長・国見裕久農学府長の下で開始。

# 農学系分野のゲノム解析の特徴

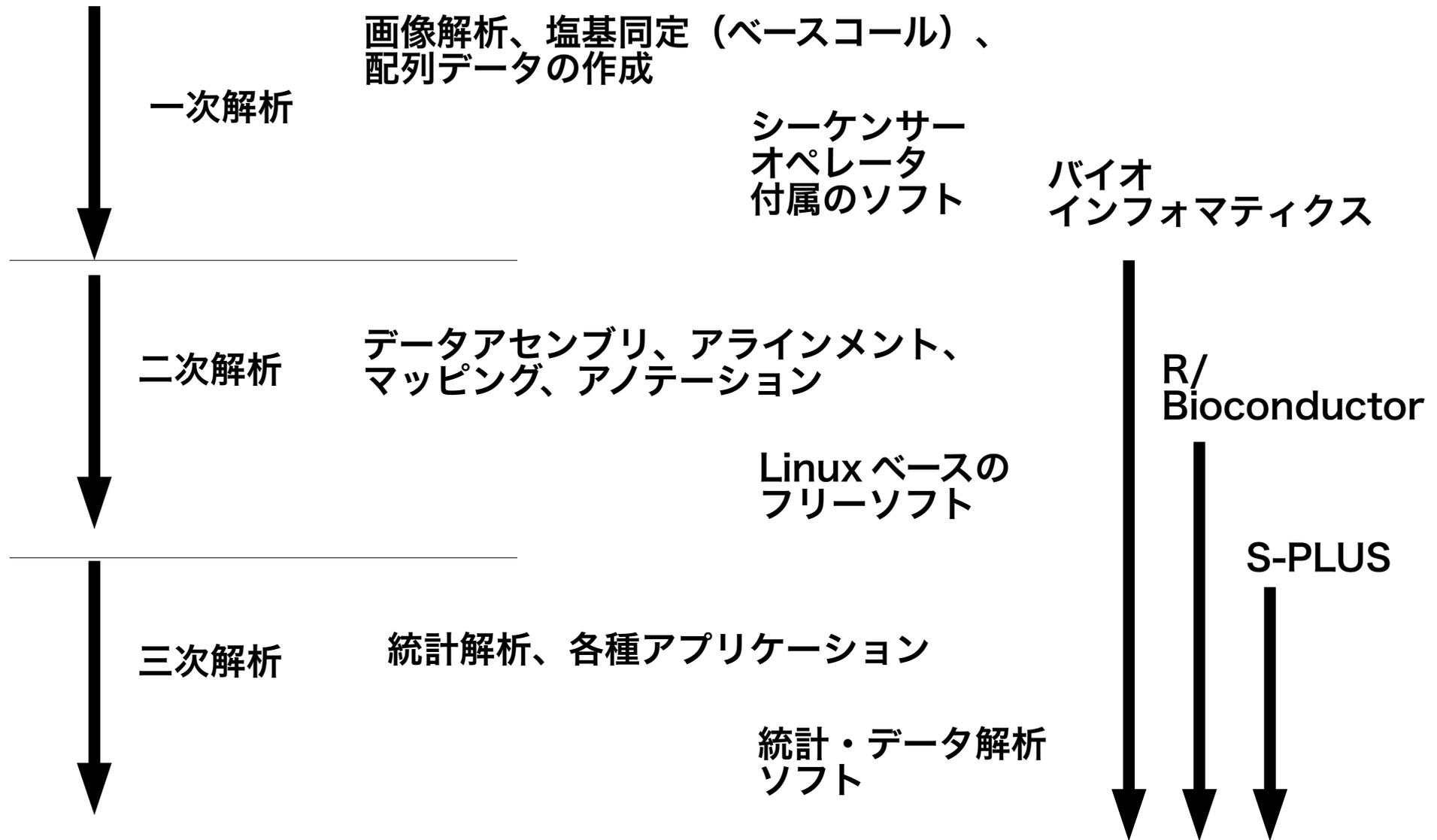
- イネ、シロイヌナズナ、ショウジョウバエなど特定の動物を除きゲノム解析がほとんどなされていない。参照配列がない。あったとしても、注釈がないモノが多い。
- 環境微生物、コケ、農作物（タバコ、ダイズ、キャベツなど）、昆虫（ダニなどを含む）など生物種が非常に多彩。
- 解析としては、de novo RNA-Seq、メタゲノム、メタRNA-Seq、de novo メタ RNA-Seq など、参照配列のないChIP-Seqなど、非常に困難な解析が多い。
- ヒトゲノムと異なり、産業上のニーズはあるものの、癌などの医療分野など巨額な研究資金を得るものが困難なものも少なくなく、限られた解析データで、結果を求められる。

# 農学系分野のゲノム解析の特徴

- 元々、日本自体のゲノム解析の一般的理解はあるとは言えず、諸外国（欧米、中国など）に比べて研究が遅れている。
- そのなかでも、農学系分野は、ヒトゲノム・医療・製薬分野に比べてもさらに遅れている。
- 教育という面では、ほとんど目立ったアクティビティに乏しい。
- 東京農工大学「農学系ゲノム科学領域における人材育成プログラム」は、学府・研究科・専攻・講座・研究教育分野・個別研究室の枠を越えた大学院教育システムを構築し、学問分野だけでなく産業分野においても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る。

# 次世代シーケンサーによるデータ解析

# 次世代シーケンサーによるデータ解析 のワークフロー



# 次世代シーケンサーデータ解析用の フリーソフトウェア

## 二次解析

QC : FastQC

トリミング : cutadapt

データアセンブリ : Velvet, Abyss, SOAPdenovo, ALLPATH2  
Oases, Trinity

マッピング : Maq, BWA, Bowtie, TopHat, ERANGE, Cufflinks, BFAST

アラインメント : Mauve, MUMmer

スプライシング : SOAPsplice

各種フォーマット用ツール : FASTX-Toolkit, SAMtools, SRA Toolkit  
vcftools, tabix

ビューアー : Tablet

アノテーション : BLAST

<http://seqanswers.com/wiki/Software/list>

[https://wiki.nbic.nl/index.php/NGS\\_Tools](https://wiki.nbic.nl/index.php/NGS_Tools)

## 三次解析

ChIP-Seq : MACS, QuEST

統計解析 : R/S-PLUS, Octave/MATLAB

# S-PLUS-ESS-EMACS/R-ESS-EMACS on Linux

The screenshot displays a Linux desktop environment with a terminal window and an Emacs editor window. The terminal window shows the execution of various R/ESS commands in a shell environment, including file listing, head, and read.table operations. The Emacs window shows the R/ESS interface with version information and R code being executed.

```
Locus_16149_Tra Locus_16149_Transcript_1/1_Confidence_1.000_Length_235 Locus_3711_Transcript_1/1_Confidence_1.000_Length_256
/media/Transcend $ wc -l result_out
9954 result_out
/media/Transcend $ ls
2result-1.1.copy.txt 2result-1.2.copy.txt~ 2result-1copy.txt Screen shot 2011-10-06 at 11.55.01 AM.png Semir
2result-1.2.copy.txt 2result-1Uniq.txt BiTC_B1e7_Asai.pdf Screen.png init.
/media/Transcend $ ls 2result-1.2.copy.txt
2result-1.2.copy.txt
/media/Transcend $ head 2result-1.2.copy.txt
Locus_1_Transcr Locus_1_Transcript_5/6_Confidence_0.061_Length_163 Locus_4678_Transcript_1/1_Confidence_1.000_Length_235
Locus_1_Transcr Locus_1_Transcript_6/6_Confidence_0.049_Length_188 Locus_4678_Transcript_1/1_Confidence_1.000_Length_235
Locus_2_Transcr Locus_2_Transcript_35/39_Confidence_0.045_Length_308 Locus_529_Transcript_1/1_Confidence_1.000_Length_430
Locus_2_Transcr Locus_2_Transcript_36/39_Confidence_0.046_Length_358 Locus_529_Transcript_1/1_Confidence_1.000_Length_430
Locus_2_Transcr Locus_2_Transcript_37/39_Confidence_0.034_Length_171 Locus_529_Transcript_1/1_Confidence_1.000_Length_430
Locus_2_Transcr Locus_2_Transcript_38/39_Confidence_0.043_Length_315 Locus_529_Transcript_1/1_Confidence_1.000_Length_430
Locus_2_Transcr Locus_2_Transcript_39/39_Confidence_0.014_Length_131 Locus_529_Transcript_1/1_Confidence_1.000_Length_430
Locus_4_Transcr Locus_4_Transcript_1/1_Confidence_1.000_Length_727 Locus_1194_Transcript_1/1_Confidence_1.000_Length_256
Locus_5_Transcr Locus_5_Transcript_1/1_Confidence_1.000_Length_1088 Locus_2324_Transcript_1/1_Confidence_1.000_Length_256
Locus_9_Transcr Locus_9_Transcript_12/18_Confidence_0.072_Length_179 Locus_529_Transcript_1/1_Confidence_1.000_Length_430
/media/Transcend $ wc -l 2result-1.2.copy.txt
11887 2result-1.2.copy.txt
/media/Transcend $ ls
2result-1.1.copy.txt 2result-1.2.copy.txt~ 2result-1copy.txt Screen shot 2011-10-06 at 11.55.01 AM.png Semir
2result-1.2.copy.txt 2result-1Uniq.txt BiTC_B1e7_Asai.pdf Screen.png init.
/media/Transcend $ head result_out
V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13
Locus_17333_Tra Locus_17333_Transcript_1/1_Confidence_1.000_Length_162 Locus_9919_Transcript_1/1_Confidence_1.000_Length_256
Locus_8676_Tra Locus_8676_Transcript_1/1_Confidence_1.000_Length_460 Locus_6063_Transcript_1/1_Confidence_1.000_Length_256
Locus_6565_Tra Locus_6565_Transcript_1/1_Confidence_1.000_Length_452 Locus_5355_Transcript_1/1_Confidence_1.000_Length_256
Locus_3899_Tra Locus_3899_Transcript_1/1_Confidence_1.000_Length_1264 Locus_6634_Transcript_1/1_Confidence_1.000_Length_256
Locus_7284_Tra Locus_7284_Transcript_1/1_Confidence_1.000_Length_760 Locus_496_Transcript_1/3_Confidence_0.267
Locus_3345_Tra Locus_3345data <- read.table(Inf, sep="\t", quote="")xx_Transcript_3/3_Confidence_0.267
Locus_2483_Tra Locus_2483_Transcript_2/6_Confidence_0.162_Length_212 Locus_529_Transcript_1/1_Confidence_1.000_Length_256
Locus_14578_Tra Locus_14578_Transcript_1/1_Confidence_1.000_Length_274 Locus_2142_Transcript_2/2_Confidence_1.000_Length_256
Locus_5322_Tra Locus_5322_Transcript_1/1_Confidence_1.000_Length_336 Locus_219_Transcript_1/3_Confidence_1.000_Length_256
/media/Transcend $ wc -l result_out
9955 result_out
/media/Transcend $ emacs result_out
/media/Transcend $ head result_out
Locus_17333_Tra Locus_17333_Transcript_1/1_Confidence_1.000_Length_162 Locus_9919_Transcript_1/1_Confidence_1.000_Length_256
Locus_8676_Tra data <- read.table(Inf, sep="\t", quote="") Locus_8676_Transcript_1/1_Confidence_1.000_Length_460
Locus_6565_Tra Locus_6565_Transcript_1/1_Confidence_1.000_Length_452 Locus_5355_Transcript_1/1_Confidence_1.000_Length_256
Locus_3899_Tra Locus_3899_Transcript_1/1_Confidence_1.000_Length_1264 Locus_6634_Transcript_1/1_Confidence_1.000_Length_256
Locus_7284_Tra Locus_7284_Transcript_1/1_Confidence_1.000_Length_760 Locus_496_Transcript_1/3_Confidence_0.267
Locus_3345_Tra Locus_3345_Transcript_3/3_Confidence_0.267 Length_161 Locus_529_Transcript_1/1_Confidence_1.000_Length_256
Locus_2483_Tra Locus_2483_Transcript_2/6_Confidence_0.162_Length_212 Locus_529_Transcript_1/1_Confidence_1.000_Length_256
Locus_14578_Tra Locus_14578_Transcript_1/1_Confidence_1.000_Length_274 Locus_2142_Transcript_2/2_Confidence_1.000_Length_256
Locus_5322_Tra Locus_5322_Transcript_1/1_Confidence_1.000_Length_336 Locus_219_Transcript_1/3_Confidence_1.000_Length_256
Locus_16149_Tra Locus_16149_Transcript_1/1_Confidence_1.000_Length_235 Locus_3711_Transcript_1/1_Confidence_1.000_Length_256
/media/Transcend $ Locus_8676_Tra Locus_8676_Transcript_1/1_Confidence_1.000_Length_460 Locus_6063_Tra Locus_6063_Transcript_1/1_Confidence_1.000_Length_256
Locus_8676_Tra: command not found
/media/Transcend $
```

```
R version 2.13.1 (2011-07-08)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-redhat-linux-gnu (64-bit)

Rは、自由なソフトウェアであり、「完全に無保証」です。
一定の条件に従えば、自由にこれを再配布することができます。
配布条件の詳細に関しては、'license()'あるいは'licence()'と入力してください。

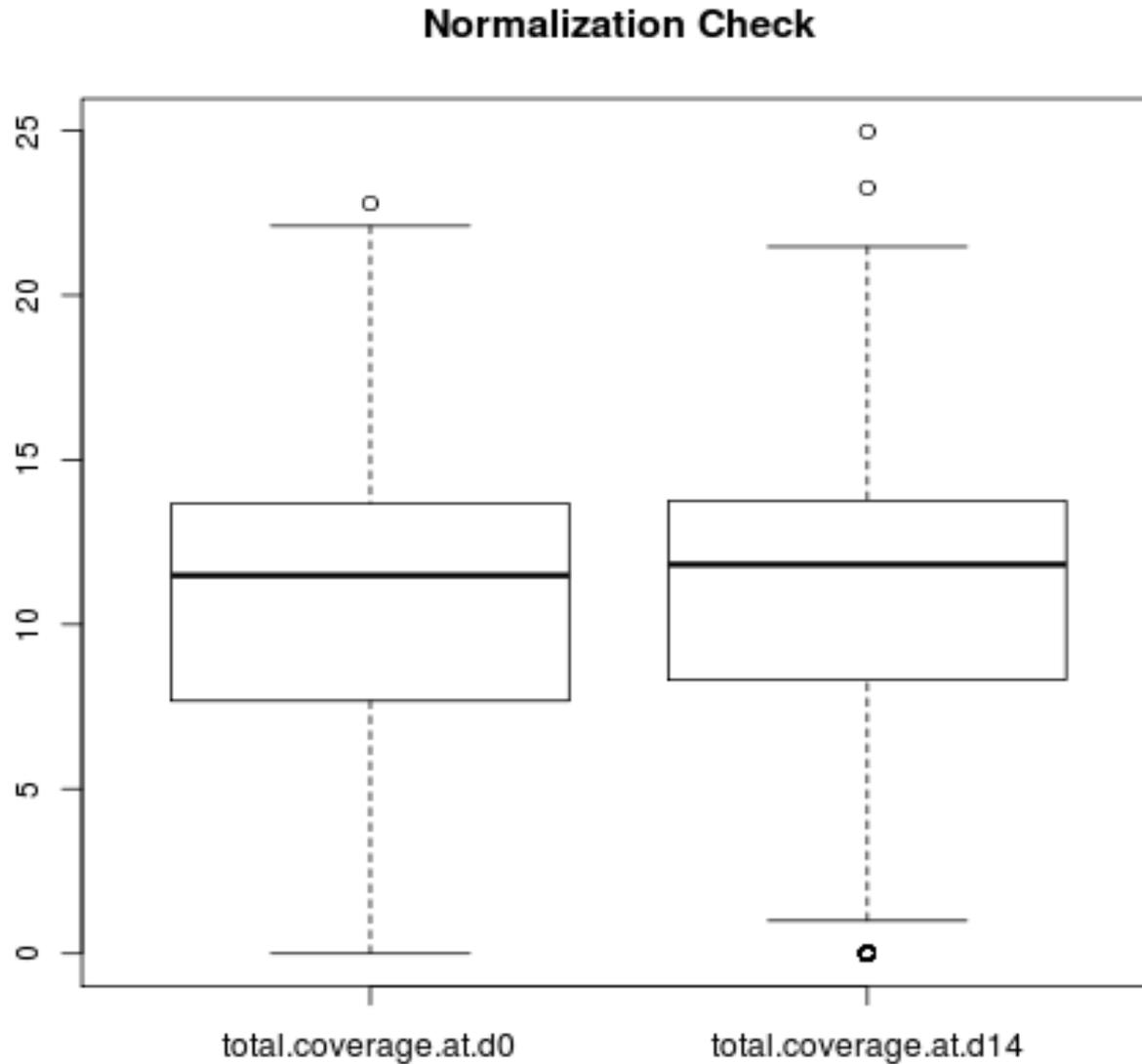
Rは多くの貢献者による共同プロジェクトです。
詳しくは'contributors()'と入力してください。
また、RやRのパッケージを出版物で引用する際の形式については
'citation()'と入力してください。

'demo()'と入力すればデモをみることができます。
'help()'とすればオンラインヘルプが出ます。
'help.start()'でHTMLブラウザによるヘルプがみられます。
'q()'と入力すればRを終了します。

[以前にセーブされたワークスペースを復元します]

> .help.ESS <- help
> options(STEM="iESS", editor="emacsclient")
> inf <- "2result-1.2.copy.txt"
> data <- read.table(inf, sep="\t", quote="")
> attach(data)
> data[order(V1, -V12),][1:100,]
      V1 V2
1 Locus_1_Transcr Locus_1_Transcript_5/6_Confidence_0.061_Length_163
2 Locus_1_Transcr Locus_1_Transcript_6/6_Confidence_0.049_Length_188
125 Locus_100_Trans Locus_100_Transcript_1/1_Confidence_1.000_Length_430
904 Locus_1000_Tra Locus_1000_Transcript_1/1_Confidence_1.000_Length_367
6754 Locus_10001_Tra Locus_10001_Transcript_1/1_Confidence_1.000_Length_495
6755 Locus_10003_Tra Locus_10003_Transcript_1/1_Confidence_1.000_Length_370
6756 Locus_10007_Tra Locus_10007_Transcript_1/1_Confidence_1.000_Length_629
6757 Locus_10008_Tra Locus_10008_Transcript_1/1_Confidence_1.000_Length_407
906 Locus_1001_Tra Locus_1001_Transcript_2/2_Confidence_1.000_Length_459
905 Locus_1001_Tra Locus_1001_Transcript_1/2_Confidence_1.000_Length_660
6758 Locus_10010_Tra Locus_10010_Transcript_1/1_Confidence_1.000_Length_473
6759 Locus_10011_Tra Locus_10011_Transcript_1/1_Confidence_1.000_Length_1291
6760 Locus_10013_Tra Locus_10013_Transcript_1/1_Confidence_1.000_Length_412
6761 Locus_10017_Tra Locus_10017_Transcript_1/1_Confidence_1.000_Length_378
6762 Locus_10018_Tra Locus_10018_Transcript_1/3_Confidence_0.259_Length_124
6763 Locus_10018_Tra Locus_10018_Transcript_2/3_Confidence_0.185_Length_123
6764 Locus_10018_Tra Locus_10018_Transcript_3/3_Confidence_0.296_Length_136
6765 Locus_10019_Tra Locus_10019_Transcript_1/1_Confidence_1.000_Length_313
6766 Locus_10021_Tra Locus_10021_Transcript_1/1_Confidence_1.000_Length_384
6767 Locus_10023_Tra Locus_10023_Transcript_1/1_Confidence_1.000_Length_426
6768 Locus_10024_Tra Locus_10024_Transcript_1/1_Confidence_1.000_Length_367
6769 Locus_10027_Tra Locus_10027_Transcript_1/1_Confidence_1.000_Length_256
6770 Locus_10029_Tra Locus_10029_Transcript_1/1_Confidence_1.000_Length_218
```

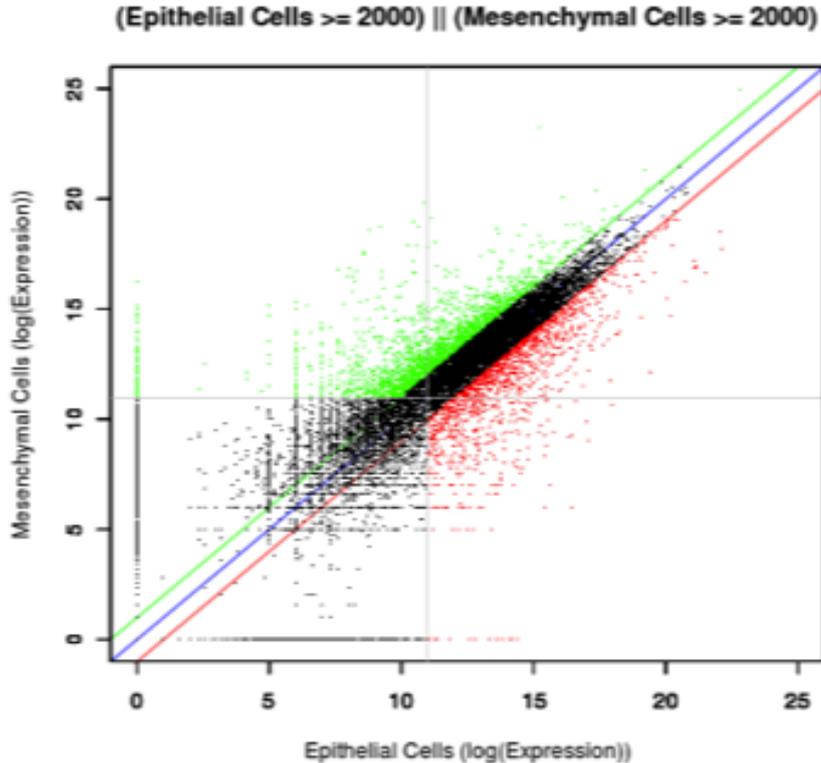
# S-PLUS/R による RNA-Seq 解析



# S-PLUS/R による RNA-Seq 解析

Mesenchymal vs. Epithelial

( $\geq 2000$  or  $\geq 2000$ )



Total = 22160

Green;  $\geq 2$  fold

n = 2690

Black;  $< 2$  fold &  $> 0.5$  fold

n = 17870

Red;  $\leq 0.5$  fold

n = 1600

# S-PLUS と R の違い

## S-PLUS

- 商用
- S+Arrayanalyzer モジュールがあったが今は開発が進んでいない
- GUI が充実
- 統計解析環境が充実
- Bioconductor の環境移植が課題

## R

- フリー
- Bioconductor パッケージあり
- コマンドベース
- 統計解析環境が充実

# BioConductor パッケージ



Search:

[Home](#)

[Install](#)

[Help](#)

[Developers](#)

[About](#)

## About Bioconductor

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, [516 software packages](#), and an active user community.

## Use Bioconductor for...

### ➤ [Microarrays](#)

Import Affymetrix, Illumina, Nimblegen, Agilent, and other platforms. Perform quality assessment, normalization, differential expression, clustering, classification, gene set enrichment, genetical genomics and other workflows for expression, exon, copy number, SNP, methylation and other assays. Access GEO, ArrayExpress, Biomart, UCSC, and other community resources.

### ➤ [High Throughput Assays](#)

Import, transform, edit, analyze and visualize flow cytometric, mass spec, HTqPCR, cell-based, and other assays.

### ➤ [Sequence Data](#)

Import fasta, fastq, ELAND, MAQ, BWA, Bowtie, BAM, gff, bed, wig, and other sequence formats. Trim, transform, align, and manipulate sequences. Perform quality assessment, ChIP-seq, differential expression, RNA-seq, and other workflows. Access the Sequence Read Archive.

### ➤ [Annotation](#)

Use microarray probe, gene, pathway, gene ontology, homology and other annotations. Access GO, KEGG, NCBI, Biomart, UCSC, vendor, and other sources.



[Mailing Lists](#)

Subscribe »



[Events](#)



[News](#)

[Re: subsetting IntegerList by list names](#)

[Advanced R Programming](#)

28 - 29 November 2011 — Heidelberg, Germany

[Bioconductor 2.9 released](#)

Following the usual 6-month cycle, the

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

# Bioconductor

- R のバイオインフォマティクス用パッケージ
- マイクロアレイのパッケージから開発開始
- MASS スペクトル、フローサイトメトリ
- アノテーション
- 次世代シーケンサーデータ解析
  - 主に発現解析と ChIP-Seq

# 次世代シーケンサーデータ解析用の R/Bioconductor パッケージ

次世代シーケンサーに特化したパッケージは、現時点で数十個あり、特に RNA-Seq 用と ChIP-Seq 用のパッケージが充実している。

IRanges, GenomicRanges, genomeIntervals: 特定領域 ( 染色体など ) のデータ処理。

Biostrings : アラインメント。

ShortRead, Rsamtools : ファイル入出力。QC、データの要約。

rtracklayer : UCSC ゲノムブラウザのトラックへの入出力データ。

BSgenome : ゲノムデータの処理。

GenomicFeatures : 種を超えて保存されている塩基配列データのアノテーション。

biomaRt : Biomart データベースへのアクセス。

SRadb : シーケンスリードアーカイブからのデータのクエリと検索を行う。

BayesPeak, ChIPpeakAnno, chipseq, ChIPseqR, ChIPsim, CSAR, DESeq, DiffBind, MEDIPS, mosaics, MotIV, nucleR, PICS, rGADEM : **ChIP-Seq 関連パッケージ (14)**。

ArrayExpressHTS, cqn, cummeRbund, DEGseq, DESeq, DEXSeq, EDASeq, EdgeR, gage, goseq, oneChannelGUI, rnaSeqMap, TSSi, tweeDEseq :  
**RNA-Seq 用発現解析パッケージ (14)**。

MEDIPS, Repitools : **Methyl-Seq 用発現解析パッケージ (2)**。

# 次世代シーケンサーデータ解析用の R/Bioconductor パッケージ

次世代シーケンサーに特化したパッケージは、現時点で数十個あり、特に **RNA-Seq** 用と **ChIP-Seq** 用のパッケージが充実している。

サンプルデータ

ChIP-Seq 用サンプルデータ。  
EatonEtAlChIPseq  
mosaicsExample  
yeastNagalakshmi

RNA-Seq 用サンプルデータ  
Pasilla  
tweeDEseqCountData,  
yeastRNASeq

## Bioconductor の読み込めるファイル形式

- Fasta
- Fastq
- ELAND
- MAQ
- BWA
- Bowtie
- SSOAP
- BAM
- Gff
- Bed
- wig

Bioconductor のサポートするデータ処理

- ・トリミング Trimming
- ・データ形式の変換 Transformation
- ・アラインメント Alignment

ドメイン特異的な解析により、品質チェック、ChIP-seq, 発現変動解析, RNA-seq, その他の方法が可能。

Sequence Read Archive というインターフェースを有する (SRADB パッケージ).

## 前処理:

- 品質評価: ShortRead, GenomicRanges
- リードの移動、トリミング、プライマー除去、その他特殊な処理: IRanges, ShortRead, Biostrings
- アラインメント: Biostrings, Bsgenome

## 各種解析:

- ChIP-seq: chipseq, ChIPseqR, CSAR, BayesPeak •
- Differential expression: DESeq, edgeR, baySeq •
- RNA-seq: Genominator

## アノテーション:

- 遺伝子データ: AnnotationDbi, org.\*.db, KEGG.db, GO.db, Category, GOstats
- ゲノムデータ: GenomicFeatures, ChIPpeakAnno

## 他の解析との統合:

- マイクロアレイの発現データ
- RNAseq and gene ontology / pathway, goseq
- HapMap, 1000 genomes, UCSC, Sequence Read Archive, GEO: ArrayExpress, rtracklayer, biomaRt, Rsamtools, GEOquery

# SPP

- CHIP-Seq のピーク検出用 R パッケージ
- ハーバード大学の Park らが開発
- インストール方法

R と Boost C++ library を設定しておく。

> R CMD INSTALL spp\_1.10.tar.gz

で、インストールできる。

# SPP 使用法

```
# load the library
```

```
library(spp);
```

```
# The following section shows how to initialize a  
cluster of 8 nodes for parallel processing
```

```
# see "snow" package manual for details.
```

```
library(snow)
```

```
cluster <- makeCluster(8);
```

# SPP 使用法

#Read in Eland alignment

```
chip.data <- read.eland.tags("chip.eland.file");
```

```
input.data <-
```

```
read.eland.tags("input.eland.file",max.eland.tag.length=32);
```

# 他の入力形式

MAQ: read.maqmap.tags() and

read.bin.maqmap.tags()

bowtie: read.bowtie.tags()

# SPP 使用法

```
# determine binding positions using wtd method  
  
bp <-  
find.binding.positions(signal.data=chip.data,control.data=input.data,fdr=fdr,whs=detection.window.halfsize,cluster=cluster)  
  
# output detected binding positions  
  
output.binding.results(bp,"example.binding.positions.txt");
```

# まとめ

- 次世代シーケンサーの開発、普及によりゲノム解析データを日常的に使う時代になった。
- データ解析には、画像データから配列データを  
得る一次解析、配列のアセンブリー、アライン  
メント、マッピングを行う二次解析、統計解  
析、生物学的解釈を行う三次解析がある。
- R は、二次解析の一部と三次解析、 S-PLUS  
は、三次解析を行うために用いる。
- 現在、 R では二次解析を行うために  
BioConductor が充実しているが、 S-PLUS  
は、 BioConductor に相当するものがない。
- BioConductor を S-PLUS をで使えるようにす  
ることが課題。