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

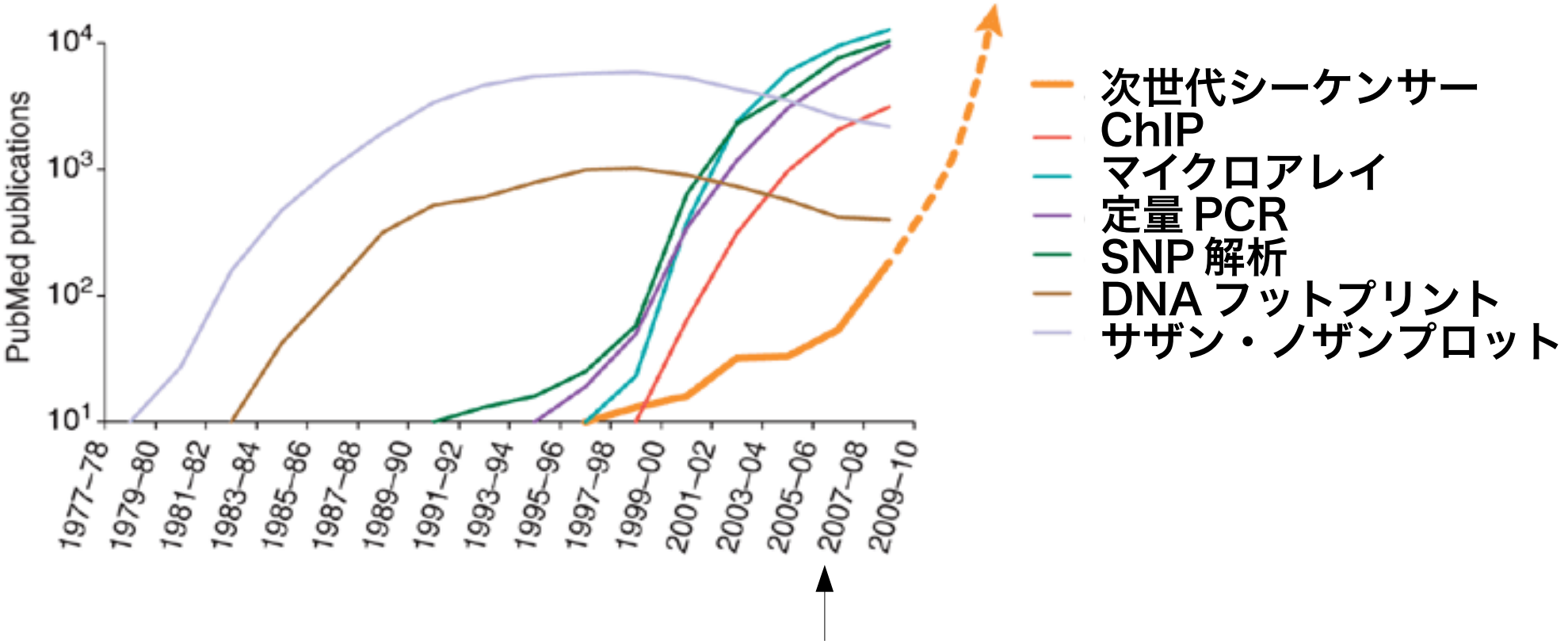
- 次世代シーケンサーとは
- 次世代シーケンサーの応用
- 農学分野におけるゲノム科学の応用とその特徴
- 次世代シーケンサーによるデータ解析
- S-PLUS/R によるデータ解析
- （デモ）今日は無理かも、、。

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

Nature Biotechnology 26, 1125 - 1133 (2008) より

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（パイロシークエンス法）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、パイロシークエ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

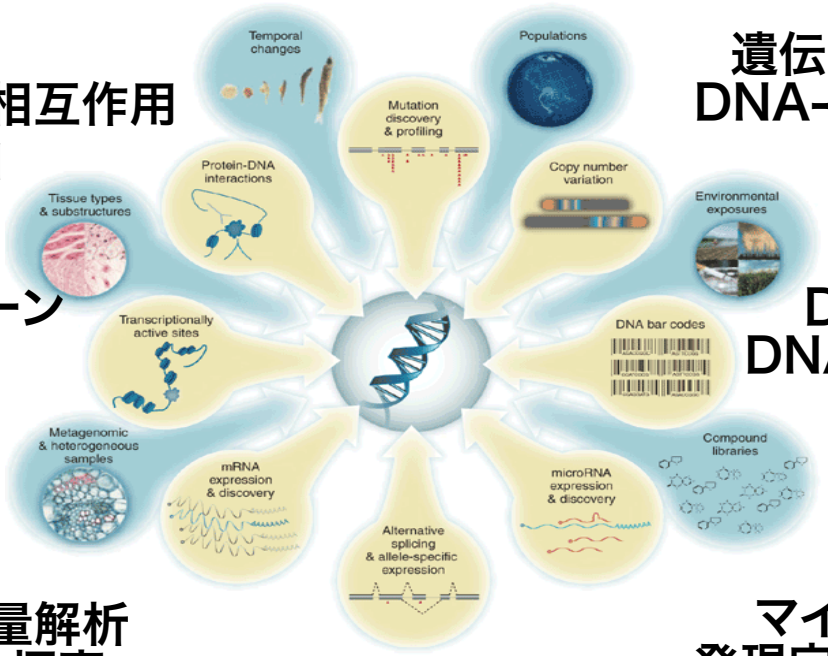
変異・多型探索
プロファイリング
DNA-Resequencing

遺伝子コピー数解析
DNA-Resequencing

タンパク質- DNA 相互作用
ChIP-Seq

クロマチン結合パターン
ChIP-Seq

DNA- バーコード
DNA-Resequencing



遺伝子発現定量解析
新規遺伝子の探索
RNA-Seq

オルタナティブ
スプライシング解析
RNA-Seq

マイクロ RNA
発現定量・探索解析
RNA-Seq

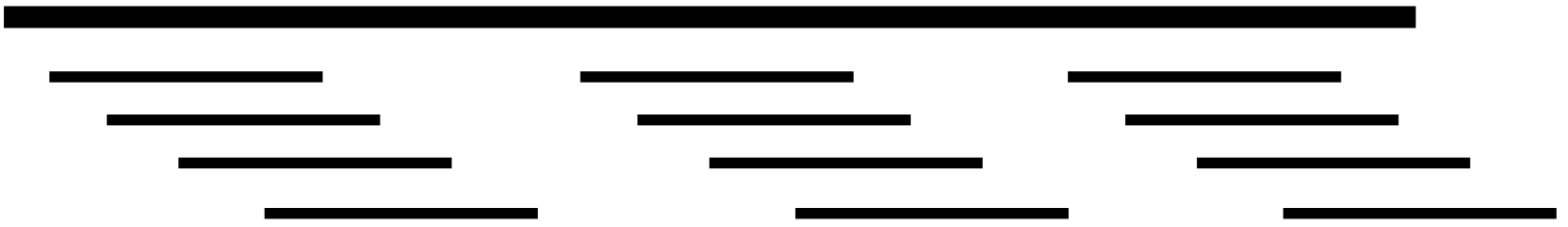
Nature Biotechnology 26, 1125 - 1133 (2008) より

3 つの主要なアプリケーション

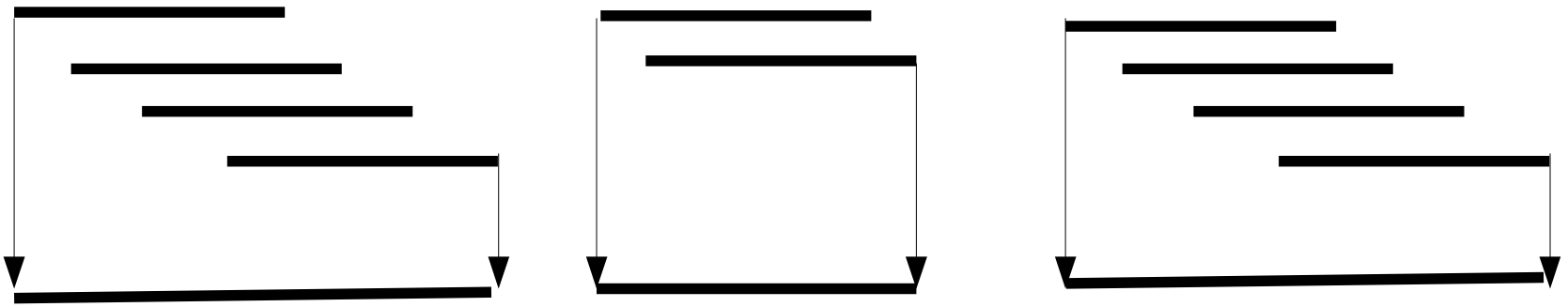
- DNA-Sequencing
  de novo Sequencing
  Resequencing (多型解析、遺伝子コピー数解析)
  メタゲノム
- RNA-Sequencing
  de novo RNA-Seq - 発現定量・新規遺伝子同定
  RNA-Seq – 発現定量・新規遺伝子同定
  メタゲノム
- タンパク質- DNA 相互作用
  ChIP-Seq - クロマチン、 RNA ポリメラーゼ、転写因子

DNA Sequencing

- マッピング　参照配列にアラインメント
  リシークエンシング・多型解析

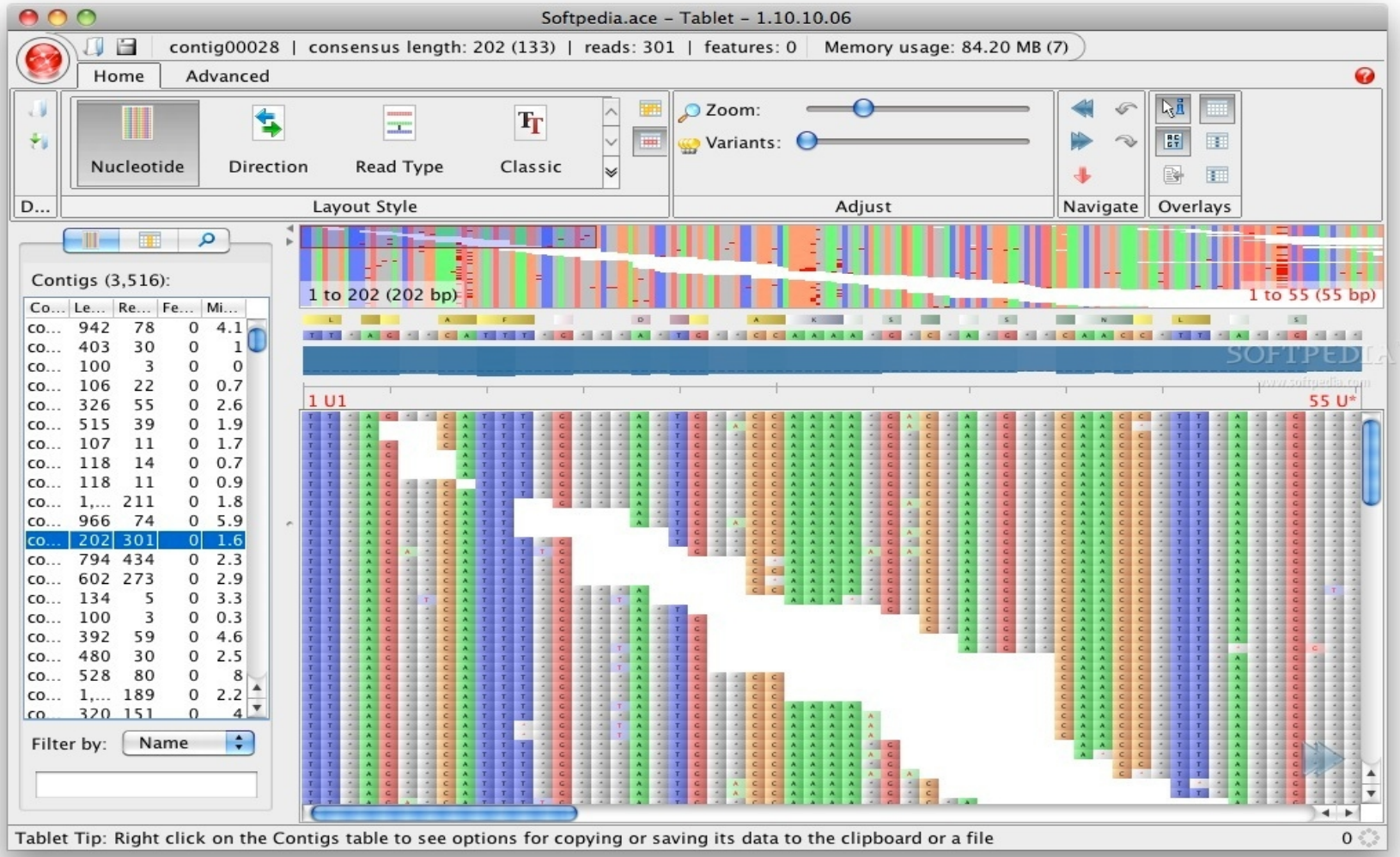


- アセンブリ　塩基配列を重ねあわせコンセンサス配列を得る　新規配列解析（de novo）



DNA Sequencing

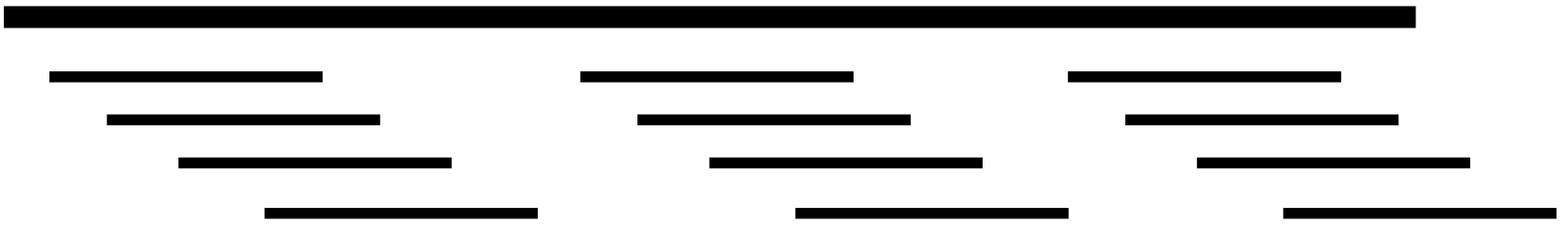
Tablet をもちいたマッピング結果の表示例



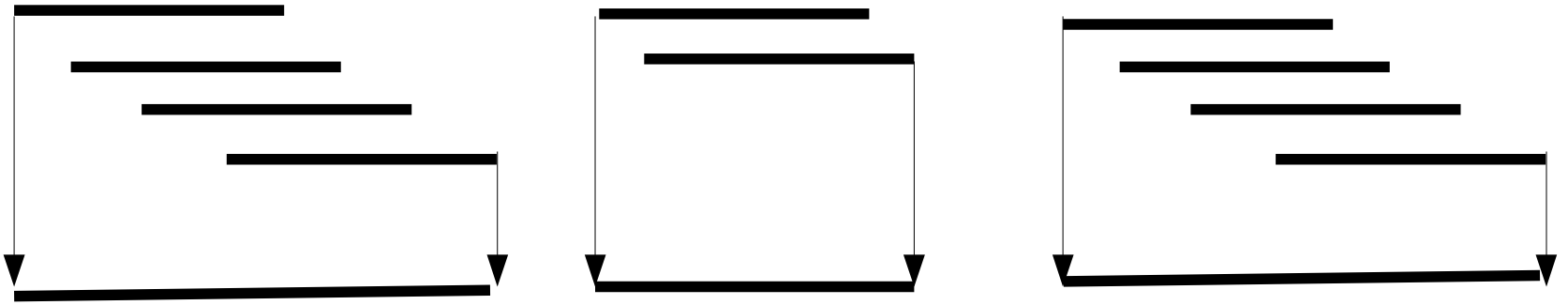
(http://mac.softpedia.com/get/Math-Scientific/)

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**

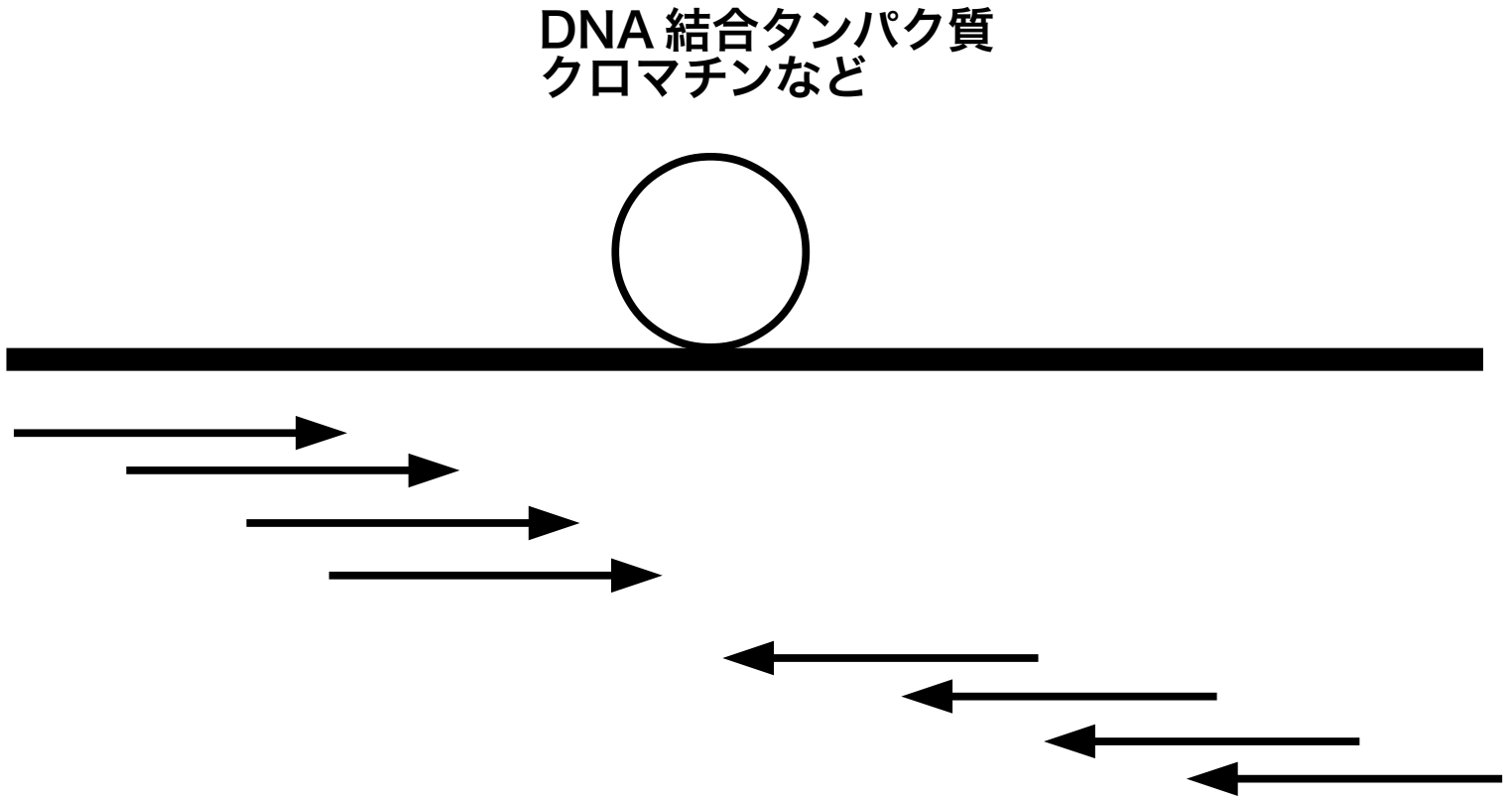


- **de novo RNA-Seq　新規遺伝子探索、発現定量解析**



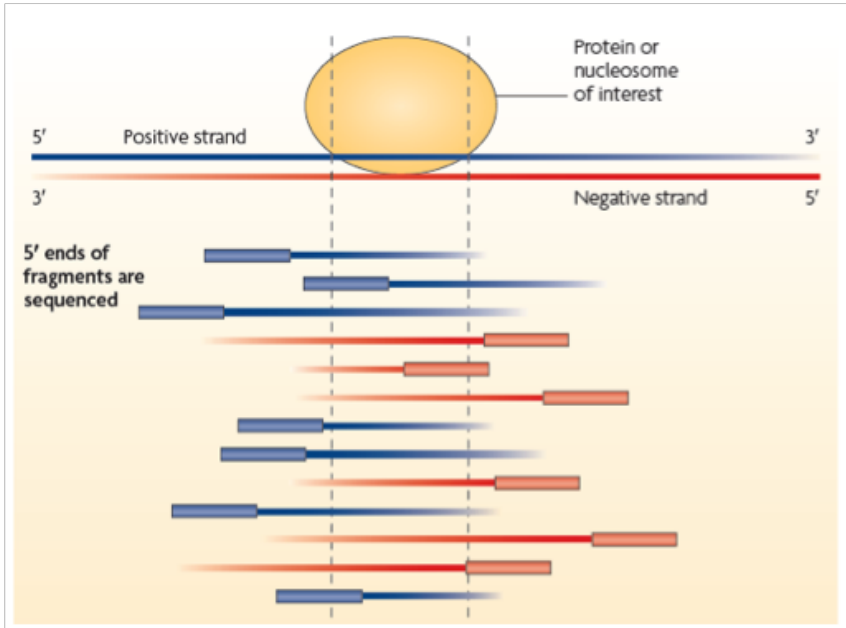
ChIP-Seq

- **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。**

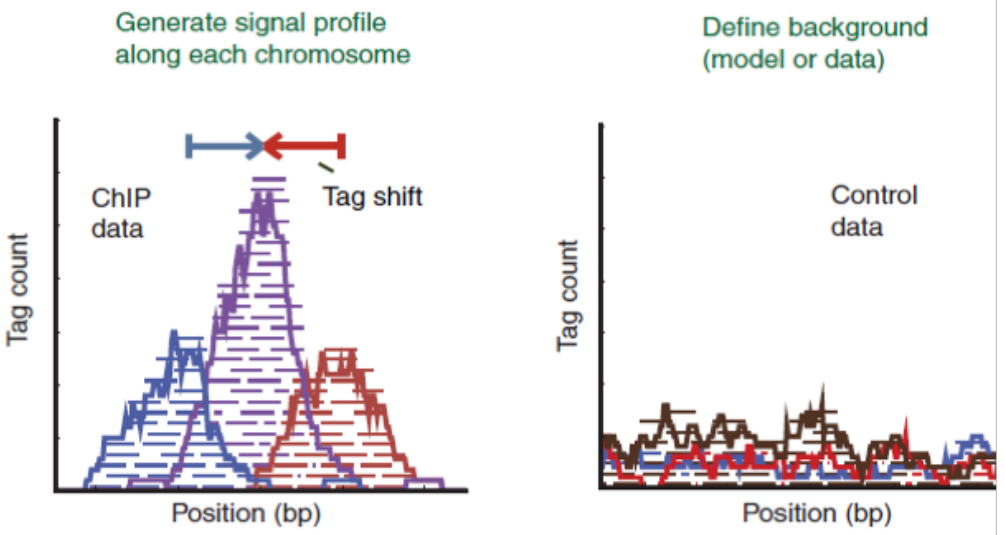


ChIP-Seq

- 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

http://www.tuat.ac.jp/~genome/

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いことが判明。
- 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、極微流量液体クトマトグラフィー - 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

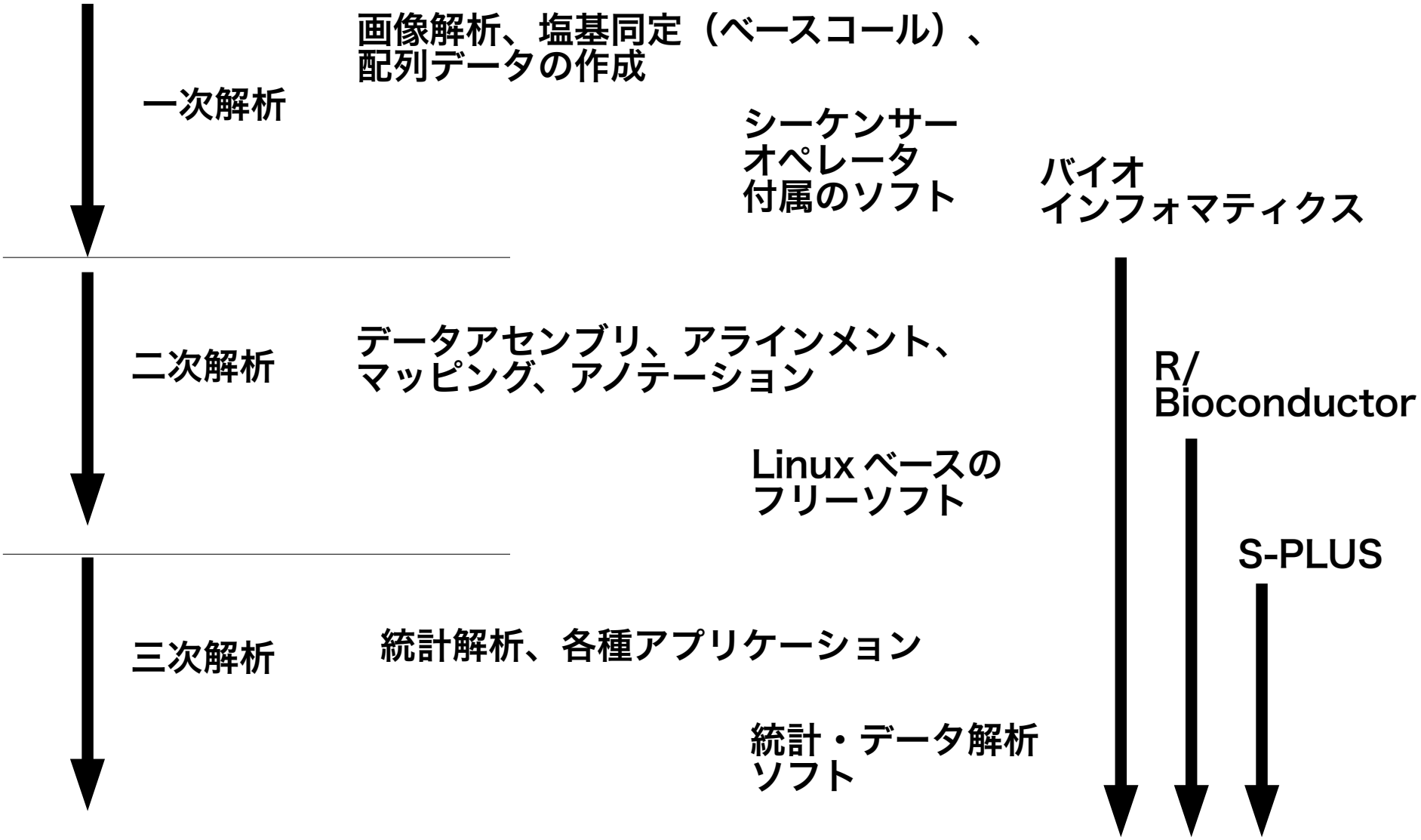
- イネ、シロイヌナズナ、ショウジョウバエなど特定の動物を除きゲノム解析がほとんどなされていない。参照配列がない。あったとしても、注釈がないモノが多い。
- 環境微生物、コケ、農作物（タバコ、ダイズ、キャベツなど）、昆虫（ダニなどを含む）など生物種が非常に多彩。
- 解析としては、 de novo RNA-Seq 、メタゲノム、メタ RNA-Seq 、 de novo メタ RNA-Seq など、参照配列のない ChIP-Seq など、非常に困難な解析が多い。
- ヒトゲノムと異なり、産業上のニーズはあるものの、癌などの医療分野など巨額な研究資金を得るものが困難なものも少なくなく、限られた解析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

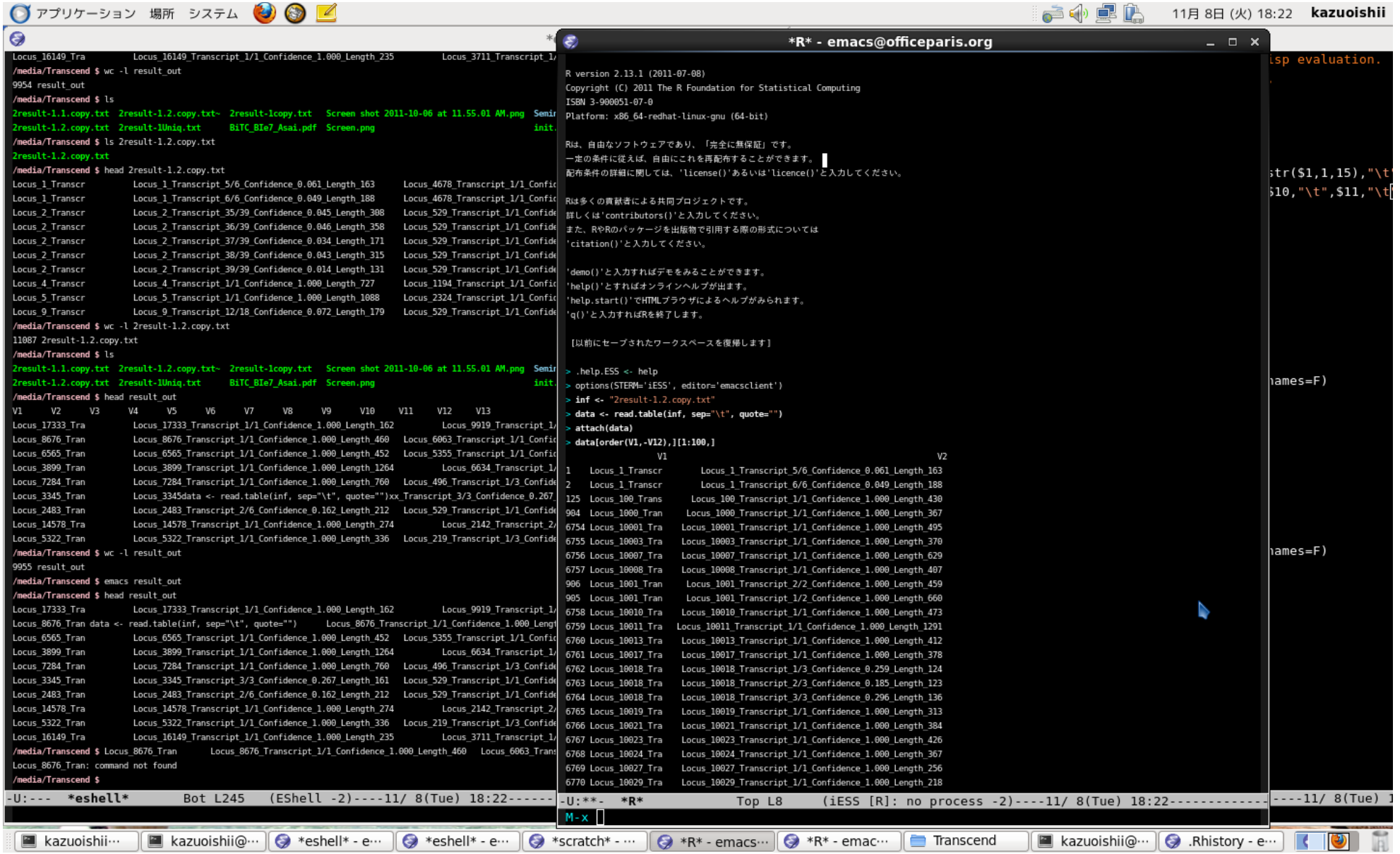
QC ： FastQC
トリミング： cutadapt
データアセンブリ： Velvet, Abyss, SOAPdenovo, ALLPATH2
Oases, Trinity
マッピング： Maq, BWA, Bowtie, TopHat, ERANGE, Cufflinks, BFAST
アラインメント： Mauve, MUMmer
スプライシング： SOAPsplice
各種フォーマット用ツール： FASTX-Toolkit, SAMtools, SRA Toolkit
vcftools, tabix
ビューアー： Tablet
アノテーション： BLAST

http://seqanswers.com/wiki/Software/list
https://wiki.nbic.nl/index.php/NGS_Tools

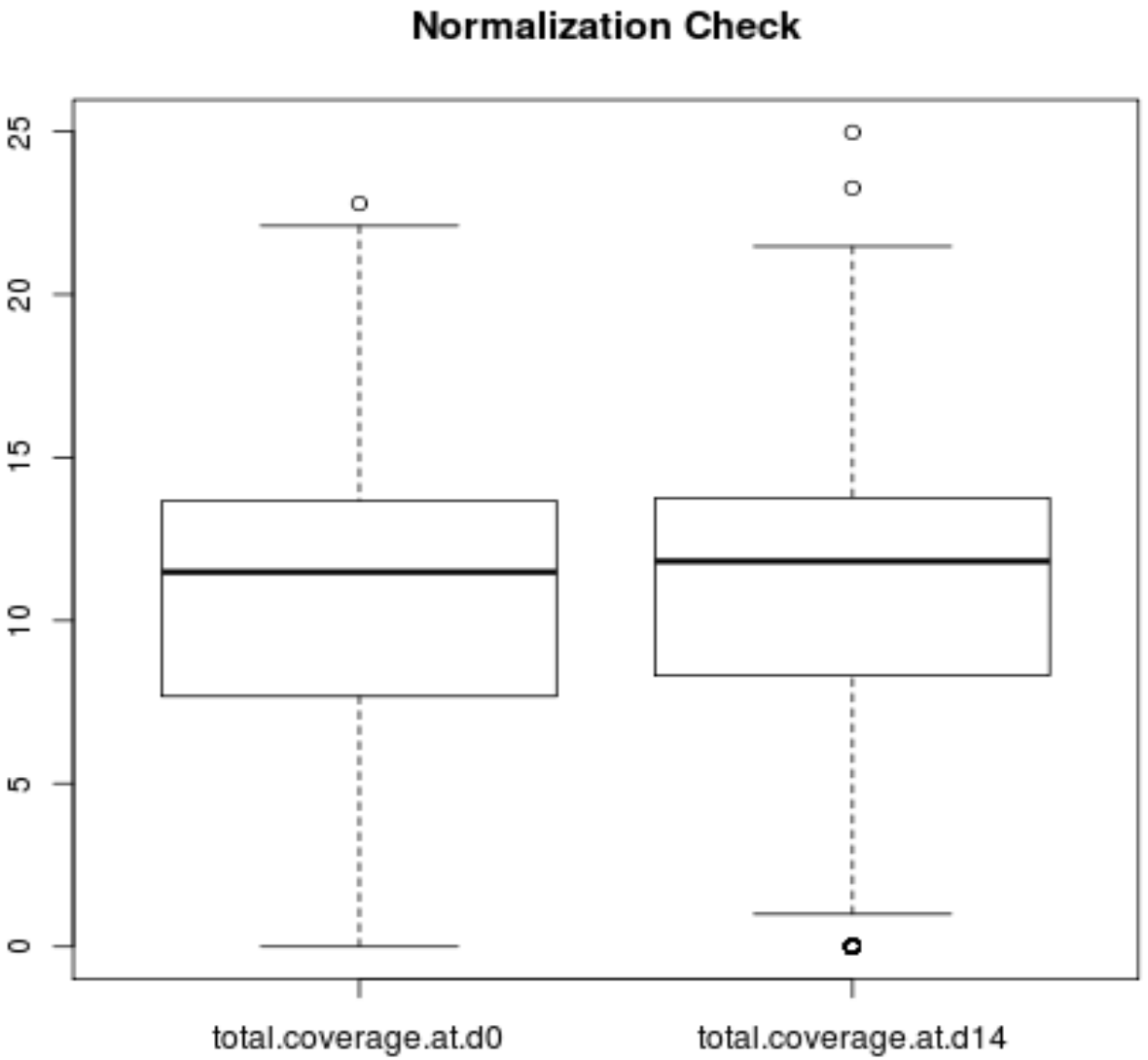
三次解析

ChIP-Seq ： MACS, QuEST
統計解析： R/S-PLUS, Octave/MATLAB

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

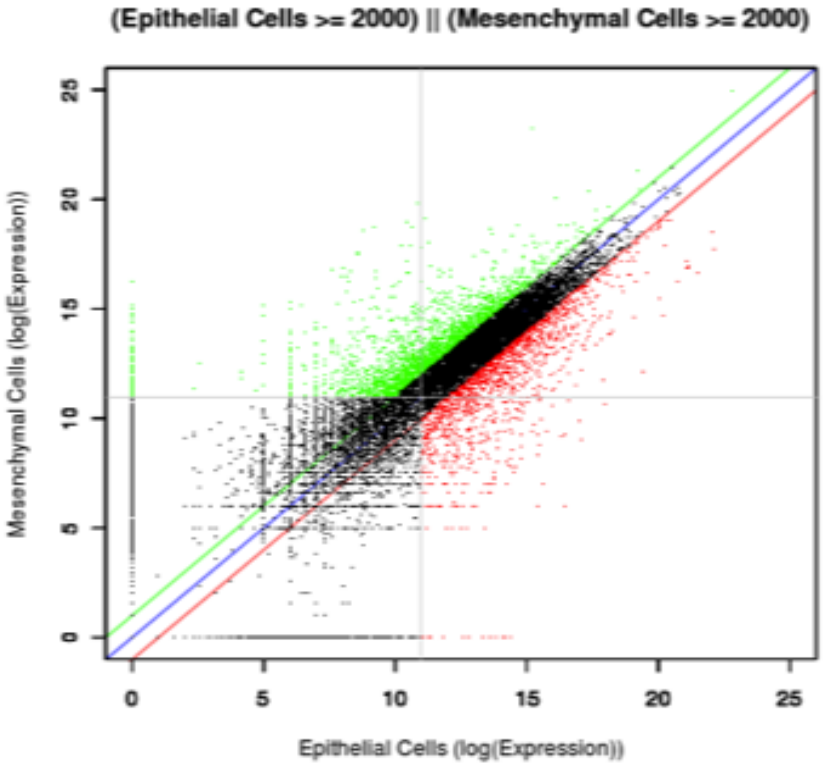


S-PLUS/R による RNA-Seq 解析



S-PLUS/R による RNA-Seq 解析

Mesenchimal vs. Epithelial
(>= 2000 or >= 2000)



Todal = 22160

Green; >= 2 fold
n = 2690

Black; < 2 fold & > 0.5 fold
n = 17870

Red; <= 0.5 fold
n = 1600

S-PLUS と R の違い

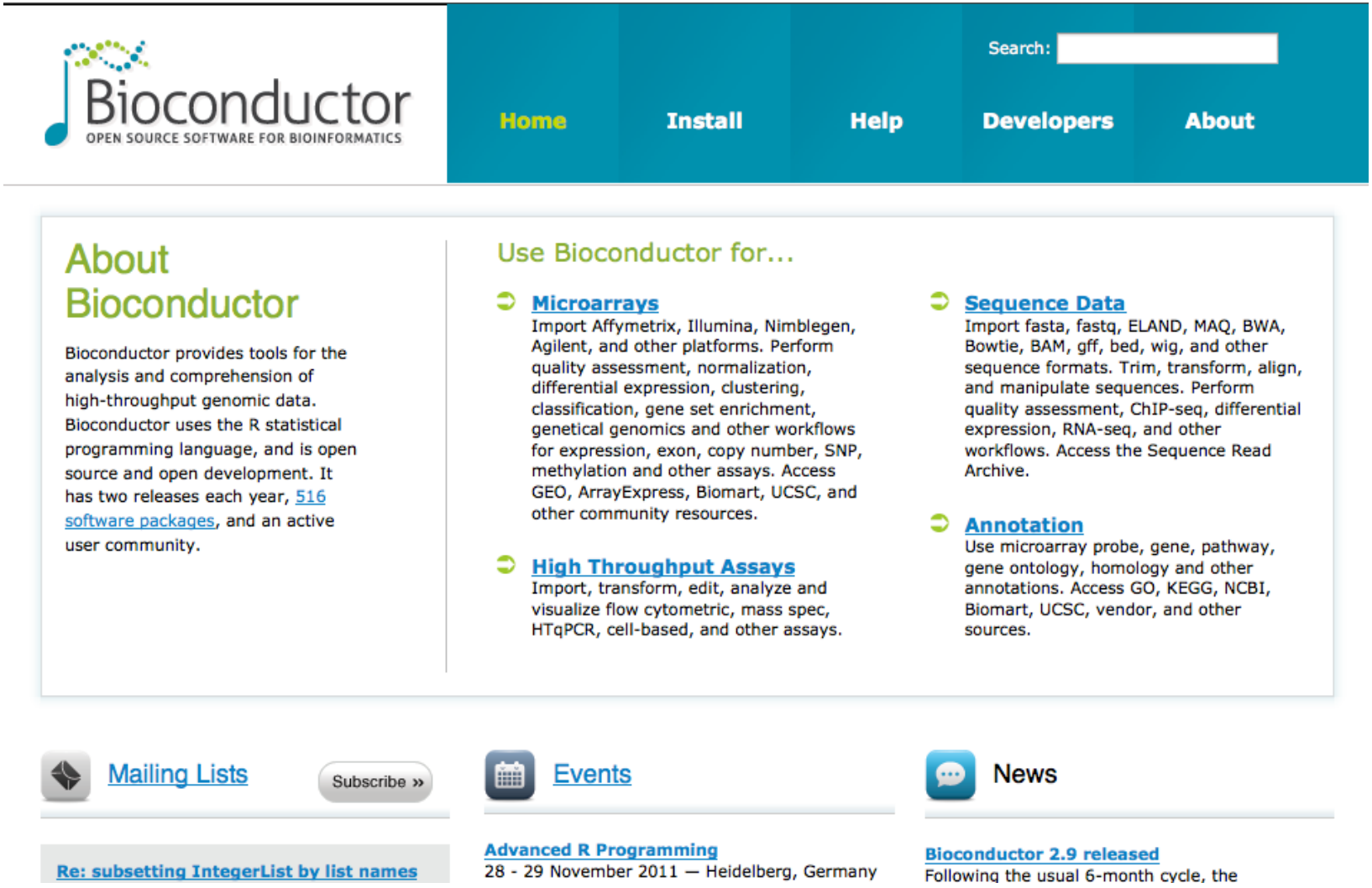
S−PLUS

- 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



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 ：ファイル出入力。 QC、データの要約。
rtracklayer ： UCSC ゲノムブラウザのトラックへの出入力データ。
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,contro
l.data=input.data,fdr=fdr,whs=detection.window.ha
lfsize,cluster=cluster)


# output detected binding positions

output.binding.results(bp,"example.binding.positio
ns.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。