トランスクリプトーム解析の現況2013（詳細版）

東京大学大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究ユニット
門田 幸二（かどた　こうじ）
http://www.iu.a.u-tokyo.ac.jp/~kadota/
kadota@iu.a.u-tokyo.ac.jp

スライドPDFはウェブから取得可能です

Google 門田 東大

http://www.iu.a.u-tokyo.ac.jp/~kadota/ 門田 幸二のホームページ

門田 幸二のホームページ

名前 門田 幸二(かどた こうじ)

所属 東京大学 大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究ユニット

身分

研究分野

所属学会

講演など(上記講義以外)(last modified: 2013.10.30) NEW

26. 題目:「Rでゲノム・トランスクリプトーム解析」, HPCIチュートリアル・バイオインフォマティクス実習コース, 生命情報工学研究センター(東京), 2014.03.07
25. 題目:「トランスクリプトーム解析の現況2013(詳細版)」, 東京大学大学院農学生命科学研究科第124回アグリバイオインフォマティクスセミナー, 東京大学(東京), 2013.11.01
24. 題目:「トランスクリプトーム解析の現況:マイクロアレイ vs. RNA-seq」, 生命医薬情報学連合大会「オミックス・計算そして創薬」・オミックス解析における実務者意見交換会, タワーホール船堀(東京), 2013.10.30
23. 題目:「食品機能解析研究とバイオインフォマティクス」, 日本農芸化学会2013年度大会・シンポジウム4SY08, 東北大学(宮城), 2013.03.27
22. 題目:「Rでトランスクリプトーム解析」, HPCIチュートリアルセミナー, 生命情報工学研究センター(東京), 2013.03.07
21. 題目:「Rでトランスクリプトーム解析」, HPCIチュートリアルセミナー, 生命情報工学研究センター(東京), 2012.03.09
20. 題目:「Rによるトランスクリプトーム解析～NGS由来塩基配列データを自在に解析する～」, Rでつなぐ次世代オミックス情報統合解析研究会, 理化学研究所横浜研究所(神奈川), 2012.02.22
19. 題目:「RNA-Seqデータ解析リテラシー」, Illumina Webinar Series・RNAシーケンスを始めよう・セッション3:データ解析, イルミナ株式会社(東京), 2011.11.17
18. 題目:「農業生物のトランスクリプトーム解析における情報処理」, 東京大学大学院農学生命科学研究科第91回アグ

研究テーマ(

トランスクリプトーム
などによって得られ
への応用を目指し
す。これまでの主な
とめにできます。ま
「(Rで)塩基配列解析

(Rで)マイクロアレイデータ解析

(last modified 2013/10/17, since 2005)

What's new?

- 2013年10月30日13:30-15:00に開
  計算　そして創薬」のフォーカスト
  いてざっくり話す予定です。(2013
- (かなり先の話ですが...)平成26年
  析を行います。情報はかなりアッ
- 「(Rで)塩基配列解析」もリニュー
- どのブラウザからでもエラーなく見
- 2013年7月18日まで公開していた
  らダウンロード可能です(88MB程
- R3.0.1がリリースされていたので
- 遺伝子セット解析の一つであるG

- はじめに (last modified 2013/07/1
- Rのインストールと起動(last modi
- Rの昔のバージョンのインストール
- 使用例(初心者向け)(last modifie
- サンプルデータ (last modified 20
- イントロ | 発現データ取得 | 公共D
- イントロ | 発現データ取得 | inSilic
- イントロ | 発現データ取得 | Array
- イントロ | 発現データ取得 | GEOc

(Rで)塩基配列解析(主に次世代シーケンサーのデータ)

(last modified 2013/10/19, since 2010)

What's new?

- 一連の解析パイプライン(RNA-seqデータ取得 -> マッピング -> カウントデータやRPKMデータ取得 -> サンプル間クラスタリングや発現変動解析およびM-A plot描画まで)をアップデートしました。項目名の一番下のほうです。(2013/10/19) NEW
- 発現変動解析用Rパッケージ TCC (ver. 1.2.0; Sun et al., BMC Bioinformatics, 2013)がBioconductorよりリリースされました。最新版を利用したい方は、R (ver. 3.0.2)をインストールしたのち、Bioconductor (ver. 2.13)をインストールしてください。10/17以降にRのインストールと起動を参考にして、通常のインストール手順で行えばそのバージョンになるはずです。(2013/10/17) NEW
- BioMart周辺の情報をアップデートしました。(2013/09/26) NEW
- 3群間比較解析(single-factorのみ)を一通り掲載しました。(2013/09/16)
- 2013年10月30日13:30-15:00に開催される、バイオインフォマティクス系学会の合同年会生命医薬情報学連合大会「オミックス・計算　そして創薬」のフォーカストセッション オミックス解析における実務者意見交換会では、トランスクリプトーム解析周辺についてざっくり話す予定です。(2013/10/08) NEW
- (かな　　アルの一部としてRでゲノム・トランスクリプトーム解析を　　013/09/11)
- どの　　しました。(2013/07/30)
- 「(R
- 201　　ジや関連ファイルはRdeennki.zipからダウンロード可能です(110MB程度)。(2013/07/30)

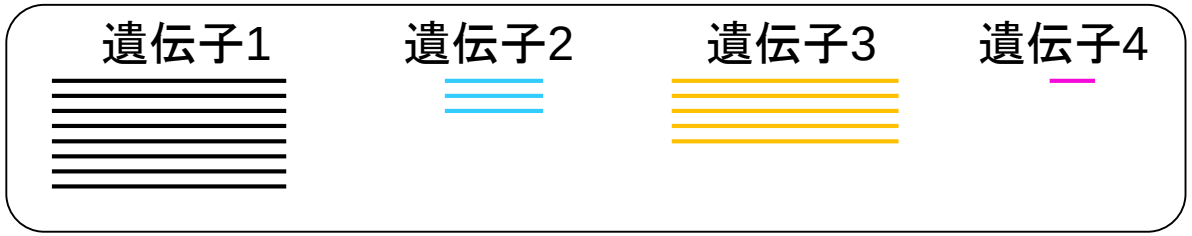
実験データ取得後のデータ解析を最小限の労力で行えるよう、2つの参考ウェブページの充実に日々取り組んでいます。

- 2013年6月6日に開催されたNAIST植物グローバル教育プロジェクト・平成25年度ワークショップのときに利用した、R(ver. 3.0.1)とTCC(ver. 1.1.99)などのインストール方法はこちら(Windows用のみ;hoge.zipはおまけ)です。

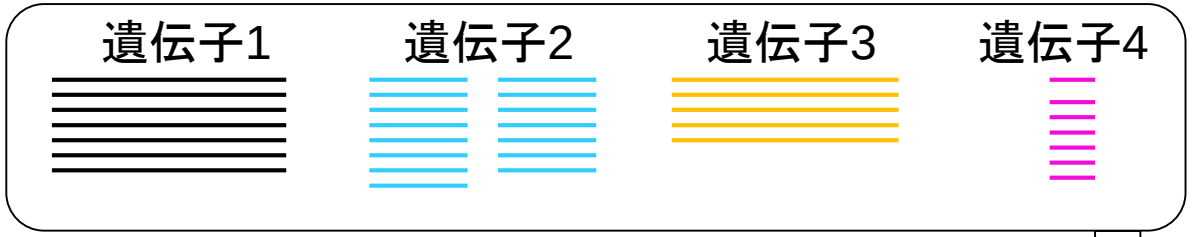
- はじめに (last modified 2013/07/30)
- Rのインストールと起動(last modified 2013/09/27) NEW
- サンプルデータ(last modified 2013/10/17) NEW
- イントロ | 一般 | ランダムに行を抽出(last modified 2013/10/10) NEW
- イントロ | 一般 | 任意の文字列を行の最初に挿入(last modified 2013/10/10) NEW

トランスクリプトーム情報を得る手段

- 光刺激前（T1）の目のトランスクリプトーム



- 光刺激後（T2）の目のトランスクリプトーム



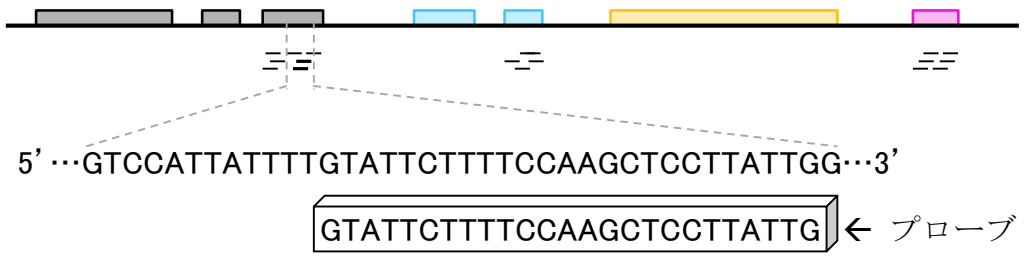
これがいわゆる
「遺伝子発現行列」

	T1	T2
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	7
…	…	…

・マイクロアレイ
・RNA-Seq
・…

ステレオタイプなイメージ

- マイクロアレイの長所
  - 取り扱いやすいデータ量（~100Mb程度）
  - 長年の実績：解析手法がほぼ確立。（Windows Rのみで解析可能）
  - 検査用チップが利用可能（MammaPrintなど）
- マイクロアレイの短所
  - 解析可能範囲が搭載転写物に限定
  - プローブが3’末端に偏っている（3’発現解析用アレイ）
  - ダイナミックレンジが狭い



ステレオタイプなイメージ

- RNA-seqの短所
  - 取り扱いづらいデータ量（数百Gb?!）
  - Windows userは自力解析が困難（ほとんどがLinux用）
  - ダイナミックレンジが広いがために?!「変」な結果に遭遇。
  - ゼロカウントデータの取り扱い（本当は気にしなくてもいいのに…）

- RNA-seqの長所
  - （多少のoff-targetは含むが）全発現転写物の解析が可能
  - 解像度：遺伝子レベル → 転写物レベル
  - ダイナミックレンジが広い

マイクロアレイ

- 機能（遺伝子セット）解析が主目的の場合にはまだ主役
  - Gene Ontology解析やパスウェイ解析
    - 実績のある市販アレイに搭載されている遺伝子のみでも「この栄養素はこのパスウェイに効いている」的な新規知見が得られればよい、という思想
    - 「個別の遺伝子の変動解析」というよりは「遺伝子セットの変動解析」
  - 同一アレイを用いている限り全体的な情報量が豊富
    - 公共データベース（GEO, ArrayExpressなど）
    - 3’発現解析用アレイが未だに使われる所以
  - 異なるアレイであっても同一生物種であればマージ可能
    - virtualArray (Heider and Alt, *BMC Bioinformatics*, **14**:75, 2013)など

NCBI

Gene Expression Omnibus

GEO Publications | FAQ | MIAME | Email GEO

NCBI » GEO » Repository browser » Platforms

Logi

登録されているサンプル数順にソートした結果

Series | Samples | Platforms | DataSets

Summary | Advance

Search

12,113 platforms

Export

Page 1 of 606 > >> Page size 20

Accession	Title	Technology	Organism(s)	Data rows	Samples	Series	Contact	Release date
GPL570	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	in situ oligonucleotide	*Homo sapiens*	54,675	87996	3198	Affymetrix, Inc.	Nov 07, 2003
GPL1261	[Mouse430_2] Affymetrix Mouse Genome 430 2.0 Array	in situ oligonucleotide	*Mus musculus*	45,101	35743	2735	Affymetrix, Inc.	May 25, 2004
GPL96	[HG-U133A] Affymetrix Human Genome U133A Array	in situ oligonucleotide	*Homo sapiens*	22,283	34187	971	Affymetrix, Inc.	Mar 11, 2002
GPL6947	Illumina HumanHT-12 V3.0 expression beadchip	oligonucleotide beads	*Homo sapiens*	49,576	16597	350	Illumina Inc.	Jun 10, 2008
GPL6244	[HuGene-1_0-st] Affymetrix Human Gene 1.0 ST Array [transcript (gene) version]	in situ oligonucleotide	*Homo sapi*					, 2007
GPL8490	Illumina HumanMethylation27 BeadChip (HumanMethylation27_270596_v.1.2)	oligonucleotide beads	*Homo sapiens*	27,578	12930	222	Illumina Inc.	Apr 27, 2009
GPL6801	[GenomeWideSNP_6] Affymetrix Genome-Wide Human SNP 6.0 Array	in situ oligonucleotide	*Homo sapiens*	1,880,794	11305	187	Affymetrix, Inc.	Apr 30, 2008
GPL6246	[MoGene-1_0-st] Affymetrix Mouse Gene 1.0 ST Array [transcript (gene) version]	in situ oligonucleotide	*Mus musculus*	35,557	11189	920	Affymetrix, Inc.	Dec 05, 2007
GPL10558	Illumina HumanHT-12 V4.0 expression beadchip	oligonucleotide beads	*Homo sapiens*	47,323	10836	399	Illumina Inc.	Jun 17, 2010
GPL6480	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F (Probe Name version)	in situ oligonucleotide	*Homo sapiens*	41,108	10283	443	Agilent Technologies	Feb 11, 2008
GPL198	[ATH1-121501] Affymetrix Arabidopsis ATH1 Genome Array	in situ oligonucleotide	*Arabidopsis thaliana*	22,810	10100	773	Affymetrix, Inc.	Jul 18, 2002
GPL1355	[Rat230_2] Affymetrix Rat Genome 230 2.0 Array	in situ oligonucleotide	*Rattus norvegicus*	31,099	10045	461	Affymetrix, Inc.	Jul 20, 2004
GPL571	[HG-U133A_2] Affymetrix Human Genome U133A 2.0 Array	in situ oligonucleotide	*Homo sapiens*	22,277	9587	417	Affymetrix, Inc.	Nov 07, 2003
GPL3718	[Mapping250K_Nsp] Affymetrix Mapping 250K Nsp SNP Array	in situ oligonucleotide	*Homo sapiens*	262,338	9522	147	Affymetrix, Inc.	May 13, 2006

Affymetrix社の3’発現解析用アレイが圧倒的

「3’発現解析用アレイ」の意味を確認

- Arabidopsis ATH1 Genome Arrayに搭載されているプローブセット”247100_at”の転写物配列（NM_126050.1）上のプローブ位置を確認

Accession	Title	Technology	Organism(s)	Data rows	Samples	Series	Contact	Release date
GPL570	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	in situ oligonucleotide	*Homo sapiens*	54,675	87996	3198	Affymetrix, Inc.	Nov 07, 2003
GPL1261	[Mouse430_2] Affymetrix Mouse Genome 430 2.0 Array	in situ oligonucleotide	*Mus musculus*	45,101	35743	2735	Affymetrix, Inc.	May 25, 2004
GPL96	[HG-U133A] Affymetrix Human Genome U133A Array	in situ oligonucleotide	*Homo sapiens*	22,283	34187	971	Affymetrix, Inc.	Mar 11, 2002
GPL6947	Illumina HumanHT-12 V3.0 expression beadchip	oligonucleotide beads	*Homo sapiens*	49,576	16597	350	Illumina Inc.	Jun 10, 2008
GPL6244	[HuGene-1_0-st] Affymetrix Human Gene 1.0 ST Array [transcript (gene) version]	in situ oligonucleotide	*Homo sapiens*	33,297	13904	742	Affymetrix, Inc.	Dec 05, 2007
GPL8490	Illumina HumanMethylation27 BeadChip (HumanMethylation27_270596_v.1.2)	oligonucleotide beads	*Homo sapiens*	27,578	12930	222	Illumina Inc.	Apr 27, 2009
GPL6801	[GenomeWideSNP_6] Affymetrix Genome-Wide Human SNP 6.0 Array	in situ oligonucleotide	*Homo sapiens*	1,880,794	11305	187	Affymetrix, Inc.	Apr 30, 2008
GPL6246	[MoGene-1_0-st] Affymetrix Mouse Gene 1.0 ST Array [transcript (gene) version]	in situ oligonucleotide	*Mus musculus*	35,557	11189	920	Affymetrix, Inc.	Dec 05, 2007
GPL10558	Illumina HumanHT-12 V4.0 expression beadchip	oligonucleotide beads	*Homo sapiens*	47,323	10836	399	Illumina Inc.	Jun 17, 2010
GPL6480	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F (Probe Name version)	in situ oligonucleotide	*Homo sapiens*	41,108	10283	443	Agilent Technologies	Feb 11, 2008
GPL198	[ATH1-121501] Affymetrix Arabidopsis ATH1 Genome Array	in situ oligonucleotide	*Arabidopsis thaliana*	22,810	10100	773	Affymetrix, Inc.	Jul 18, 2002
GPL1355	[Rat230_2] Affymetrix Rat Genome 230 2.0 Array	in situ oligonucleotide	*Rattus norvegicus*	31,099	10045	461	Affymetrix, Inc.	Jul 20, 2004
GPL571	[HG-U133A_2] Affymetrix Human Genome U133A 2.0 Array	in situ oligonucleotide	*Homo sapiens*	22,277	9587	417	Affymetrix, Inc.	Nov 07, 2003
GPL3718	[Mapping250K_Nsp] Affymetrix Mapping 250K Nsp SNP Array	in situ oligonucleotide	*Homo sapiens*	262,338	9522	147	Affymetrix, Inc.	May 13, 2006

統合遺伝子検索
GGRNA ver.2

Help | Advanced search | English 旧バージョン

Naito and Bono, *Nucleic Acids Res.*, *40*: W592-W596, 2012

247100_at 検索 Arabidopsis thaliana (thale cress)

2013-10-22 16:23:50, GGRNA : RefSeq release 61 (Sep, 2013)

Summary:

- seq:TTGCTCAAAGCCTGTCGAATTCACA (1)
- seq:TTAGCGGGAGACCGATCACACCCAG (1)
- seq:GGGACCACACACGAGTTTTTAGCGG (1)
- seq:GGGAAATGCAGTTGTGGGGACTACT (1)
- seq:GATGCTGCTTGATCTAGTTGATGAT (1)
- seq:GAGAGGCTATTGTGCATCAGCATAG (1)
- seq:GACTAAACCAGGAACCATTATTCGT (1)
- seq:GACCGATCACACCCAGAGATAGAGA (1)
- seq:GAAATTGGCTATTACATACGGGTTA (1)
- seq:AGATAGGACAAGGTTCCATCATTTC (1)
- seq:AAGTAACGAAGCTCATCTCTAAGAT (1)
- **INTERSECTION (1)**

Results:

トップ50件を表示。検索語に色がつきます。重なると色が濃く表示されます。

Arabidopsis thaliana pentatricopeptide repeat-containing protein mRNA, complete c

gcg**ttgctcaaagcctgtcgaattcaca**aaaacatcgaattgggagaggaaattggagaaatcttaattgcaatagatccatat
ggcaaatattcatgctatggataaaaagtgggacaaagcagctgaaacaagaagattgatgaaagaacaaggagtagcaaaag
ggaa**gggaccacacacgagtttttagcgggagaccgatcacacccagagatagaga**aaattcaatctaaatggagaat
aacgggtacgtaccagagttagaaga**gatgctgcttgatctagttgatgat**gatgaaa**gagaggctattgtgcatcagcat**
tacgggttaatcaa**gactaaaccaggaaccattattcgt**ataatgaagaatctccgagtatgcaaagattgtcaca**aagtaa**
caagagggatattgtaatgag**agataggacaaggttccatcatttc**agagat**gggaaatgcagttgtggggactact**gg.

position 1264 1480 1498 1507 1602 1634 1662 1692 1754 1803 1834

Synonym: K1F13.18; K1F13_18

NM_126050.1 - Arabidopsis thaliana (thale cress) - NCBI - TAIR

atgaatgtgatctcatgctccttctcattggaacataatctctacgagacaatgt
cttgtctccagagatgctcaaagcaagaagaactgaagcaaatccacgctcgcat
gctgaaaactggcttgatgcaggattcttatgcaatcacaaagtttctttctttc
tgcatttcttcaacgtcttccgacttttgccttatgcccagattgtgtttgacg
ggtttgatcgaccagatactttcttgtggaacctaatgatcagagggttctcgtg
ctcagacgaacccgagaggtctcttctcctgtatcaacgtatgctctgttcttca
gctcctcataacgcgtatacttttccgtctcttctcaaagcttgttcgaacctat
ctgcatttgaagaaacaacgcaaattcacgcacagatcacgaaacttggatatga
aaatgatgtctatgcagtgaattctctgattaattcatatgctgtgaccgggaat
ttcaagctagctcaccttctctttgacagaatccccgaacctgatgatgtctcgt
ac
ct
gggtatgttcaagcagacatgaacaaggaagctctgcaattgtttcatgaaatgc
agaattcagatgttgagcctgataatgtttccctagctaatgctctctcagcttg
tgctcagctcggagcactcgagcaagggaaatggatccattcctatcttaataag
acaagaatcagaatggactctgttttgggttgtgttcttatagatatgtatgcaa
agtgcggtgaaatggaagaagctttggaagttttcaagaatattaagaaaaaatc
agtgcaagcatggacagctttgatttcaggatatgcataccatggccatggaaga
gaagctattagcaaattcatggagatgcaaaagatgggaattaagccaaatgtga
tcactttcactgcggttctcacggcttgcagctacacaggactagttgaagaagg
aaagttgatattctacagcatggagagagattacaacctgaaaccgaccatcgag
cattatggctgtattgttgatttactcggtcgagctggattgcttgatgaagcaa
aacgtttcattcaggagatgccattgaagccaaatgctgtgatatggggtgcg**tt**
gctcaaagcctgtcgaattcacaaaaacatcgaattgggagaggaaattggagaa
atcttaattgcaatagatccatatcatggcggaagatatgttcataaggcaaata
ttcatgctatggataaaaagtgggacaaagcagctgaaacaagaagattgatgaa
agaacaaggagtagcaaaagttccaggatgtagtacaattagcttggaa**gggacc**
acacacgagtttttagcgggagaccgatcacacccagagatagagaaaattcaat
ctaaatggagaatcatgagaaggaaacttgaggaaaacgggtacgtaccagagtt
agaaga**gatgctgcttgatctagttgatgat**gatgaaa**gagaggctattgtgcat**
cagcatagcgagaaattggctattacatacgggtta**atcaagactaaaccaggaa**
ccattattcgtataatgaagaatctccgagtatgcaaagattgtcaca**aagtaac**
gaagctcatctctaagatatacaagagggatattgtaatgag**agataggacaagg**
ttccatcatttcagagat**gggaaatgcagttgtggggactact**ggtaa

「3’発現解析用アレイ」の意味がよく分かります

マイクロアレイ（デバイスの進歩）

- 3’発現解析用アレイ → exon array → transcriptome array
  - Affymetrix Human Transcriptome Array (HTA 2.0)
  - Furney et al., *Cancer Discov*., **3**: 1122-1129, 2013.
  - GPL17585(exon level)
  - GPL17586(gene level)

転写物数は有限であるため、
RNA−seqによる網羅的な同定後は、
「トランスクリプトームアレイ」に移
行するほうがお手軽かもしれない

3’発現解析用アレイ、エクソンアレイ、
HTA2.0アレイのプローブの比較の図
（どこから得たか忘れました
…Affymetrixさんから直接もらったかも）

マイクロアレイ(前処理法の進歩)

- よく使われてきた方法
  - RMA (Irizarry et al., *Biostatistics*, 2003)
    - **特徴**:データセット中の複数のアレイデータ情報を利用(multi-array basis)
    - **probe level正規化**:quantile normalization
    - **要約統計量**:median polish
  - MAS5 (Hubbell et al., *Bioinformatics*, 2002)
    - **特徴**:アレイごとに独立して前処理(正規化)を実行(per-array basis)
    - **probe level正規化**:なし
    - **要約統計量**:one-step Tukey's biweight

RMAがいいという評価がほぼ定着

マイクロアレイ(前処理法の進歩)

- RMAの問題点
  - 本当はばらつきの大きいデータを過小評価
    - median polishを利用しているため、手続き的に必要以上にサンプル間で似た結果を返す(Giorgi et al., *BMC Bioinformatics*, **11**: 553, 2010)
  - サンプル数の増減のたびに、RMA再実行の必要性
    - quantile normalizationを利用しているため、リファレンス分布が変化。例えばサンプル数の増加の場合、元々存在していたサンプルの数値も変わってしまう。
- MAS5の問題点
  - 低発現領域でばらつきが大きい傾向
    - Absent callのデータをフィルタリング(すればいいのに)しないため(McClintick and Edenberg, *BMC Bioinformatics*, **7**: 49, 2006)

このあたりを認識できていないヒト、意外に多いのかも…

マイクロアレイ（前処理法の進歩）

- fRMA(McCall et al., *Biostatistics*, **11**: 242-253, 2010)
  - RMAの改良版(サンプル数の増減の影響を受けない)
  - シグナル強度を得たいデータセット以外の多様なデータを用いて、正規化に必要な「リファレンス分布」と「プローブ効果の推定値」の情報を予め取得（パラメータをfrozenしておき、それを新規サンプルに独立に適用）

    → 目的データセット中のサンプルのシグナル強度を(データセット中の他のサンプルの影響を受けずに)得ることが可能
  - RMA → refRMA → fRMA。refRMA(Katz et al., 2006)では一定と仮定していたバッチ効果を考慮
  - 短所
    - パラメータ推定が大変らしく、Affymetrixチップの一部しか利用不可能
    - Affymetrix Exon array用のパラメータ提供が論文に…(McCall et al., 2012)

マイクロアレイ（前処理法の進歩）

- IRON(Welsh et al., *BMC Bioinformatics*, **14**: 153, 2013)
  - 解析データセットの中からリファレンスサンプルを一つ選び、それと他のサンプルをペアワイズで正規化。リファレンスが固定されているので、サンプル数の増減の影響を受けない。要約統計量はTukey's biweight。
- RMX (Kohl and Deigner, *BMC Bioinfo*rmatics, **11**: 583, 2010)
  - MAS5と同じで、要約統計量の計算部分がrobust rmx estimatorに置き換わったもの。正真正銘per-array basisのものであるため、ややこしいことを考えなくてよく、使用感もよい(個人の感想です)。
- ベイズオンライン学習(Lahti et al., *Nucleic Acids Res*., **41**: e110, 2013)
  - Affymetrix以外の様々な市販アレイにも対応した拡張性の高いアルゴリズム

進展していますね。。。

マイクロアレイ（前処理法の進歩）

- よく使われてきた方法
  - RMA (Irizarry et al., *Biostatistics*, **4**: 249-264, 2003)
    - 要約統計量：median polish
  - MAS5 (Hubbell et al., *Bioinformatics*, **18**: 1585-1592, 2002)
    - 要約統計量：one-step Tukey's biweight

RMAがいいという評価がほぼ定着しているが…

…

- 比較的最近の方法
  - fRMA(McCall et al., *Biostatistics*, **11**: 242-253, 2010)
    - 要約統計量：robust weighted average
  - RMX (Kohl and Deigner, *BMC Bioinf*ormatics, **11**: 583, 2010)
    - 要約統計量：robust rmx estimator
  - IRON(Welsh et al., *BMC Bioinformatics*, **14**: 153, 2013)
    - 要約統計量：one-step Tukey's biweight
  - ベイズオンライン学習(Lahti et al., *Nucleic Acids Res*., **41**: e110, 2013)

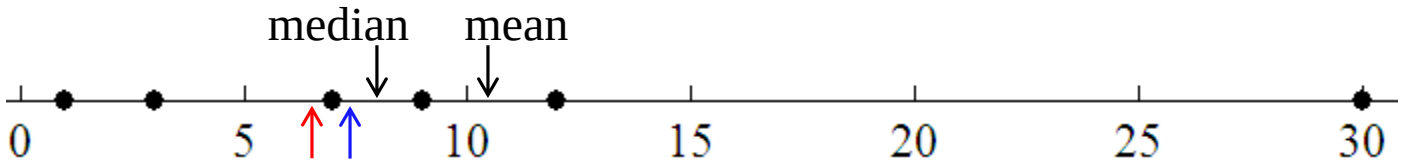
…

Tukey's biweightやRMX

- 重みつき平均の一種
  - 外れ値の影響をなるべく受けないようにしたい
    - "中央"付近の数値には1に近い重み
    - "中央"から遠く離れるほど重みを軽くしたい（0に近い重み）
  - 例：$\boldsymbol{x}$ = （1, 3, 7, 9, 12, 30）の重みつき平均（weighted mean）

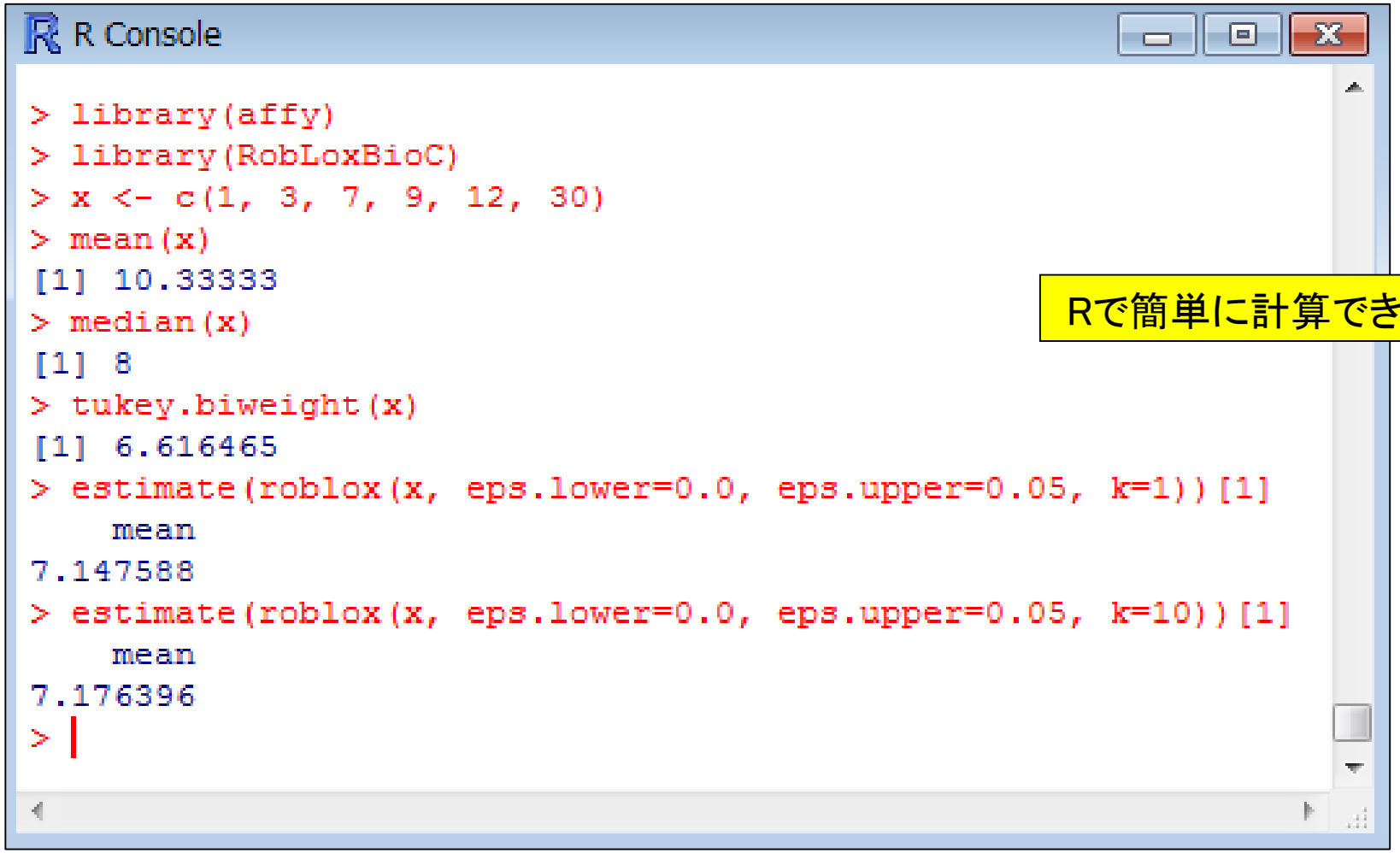
    x_1 x_2 x_3 x_4 x_5 x_6

    - mean = (1+3+7+9+12+30)/6=10.3
    - median *M*= (7+9)/2=8
    - one-step Tukey biweight = 6.61
    - one-step rmx estimate = 7.148
    - 10-step rmx estimate = 7.176



Tukey's biweightやRMX

- 例：$\boldsymbol{x}$ = (1, 3, 7, 9, 12, 30)の重みつき平均(weighted mean)



Rで簡単に計算できます

一通りのRNA-seq解析はRで可能になった

- **大人数のハンズオン講義**でLinuxはアリエナイ
  - 基本的なコマンド：cd, pwd, ls
  - 「スペース」の概念：ファイル名中に普通に存在…
    - 「bowtie -k2 filename」→「bowtie-k2fil ename」
  - エラーの認識：ごく初期段階でダメになってるのに…

Linuxコマンドは教える側も教えられる側も鬼門

今はLinuxコマンド抜きで一通り解析可能

- *SRAdb* (Zhuら, *BMC Bioinformatics*, **14**: 19, 2013)
  - 公共DBからのRNA-seqデータ（FASTQファイル）取得
- ***QuasR* (Lerchら, unpublished)**
  - **リファレンス配列（ゲノム or トランスクリプトーム）へのマッピング**
    - **Bowtie (Langmeadら, 2009) or SpliceMap (Auら, 2010)を選択可能**
    - **出力はBAM形式ファイル、QCレポートも**
  - **遺伝子アノテーション情報をもとにカウントデータ取得**
    - ***GenomicFeatures* (Lawrenceら, 2013)で得られるTranscriptDbオブジェクトを利用**
    - **UCSC known genesやEnsembl genesのカウントデータなど**
- *TCC* (Sunら, *BMC Bioinformatics*, **14**: 219, 2013)
  - 内部的に*edgeR* (Robinsonら, 2010)や*DESeq* (Anders, 2010)などを用いて頑健な発現変動解析を実行

アセンブル以外ならWindows（のR）上でどうにかなる時代がやってきました

解析例：SRP017142のデータ

- *SRAdb*を用いたgzip圧縮FASTQ形式ファイルのダウンロード
  - Neyret-Kahn et al., *Genome Res*., **23**: 1563-1579, 2013
    - 複製あり2群間比較用ヒトRNA-seqデータ(3 Ras vs. 3 Proliferative)

FileName	SampleName
SRR616151.fastq.gz	Pro_rep1
SRR616152.fastq.gz	Pro_rep2
SRR616153.fastq.gz	Pro_rep3
SRR616154.fastq.gz	Ras_rep1
SRR616155.fastq.gz	Ras_rep2
SRR616156.fastq.gz	Ras_rep3

計6GB程度。*QuasR*パッケージは圧縮ファイルのままでマッピング可能

- *QuasR* (Bowtie)を用いたヒトゲノムへのマッピング
  - *BSgenome.Hsapiens.UCSC.hg19*パッケージを利用
  - 18種類程度の生物種のゲノム配列がRパッケージとして利用可能
    - シロイヌナズナの場合：*BSgenome.Athaliana.TAIR.TAIR9*
    - ショウジョウバエの場合：*BSgenome.Dmelanogaster.UCSC.dm3*

解析例：SRP017142のデータ

- *QuasR* (Bowtie)を用いたカウント情報取得

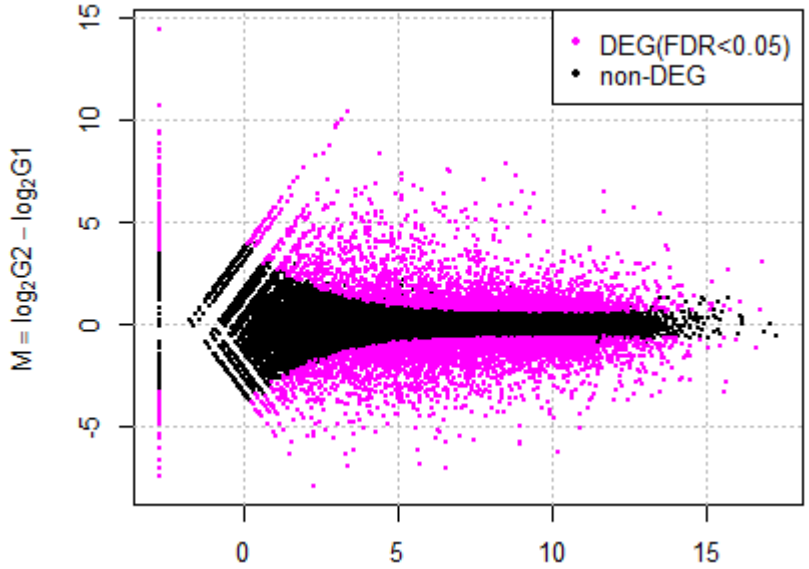
raw countデータ

	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG00000000003	480	513	366	124	271	366
ENSG00000000005	0	0	0	1	0	0
ENSG00000000419	282	354	208	165	301	209
ENSG00000000457	167	198	155	156	248	129
ENSG00000000460	114	112	101	55	81	59
…						

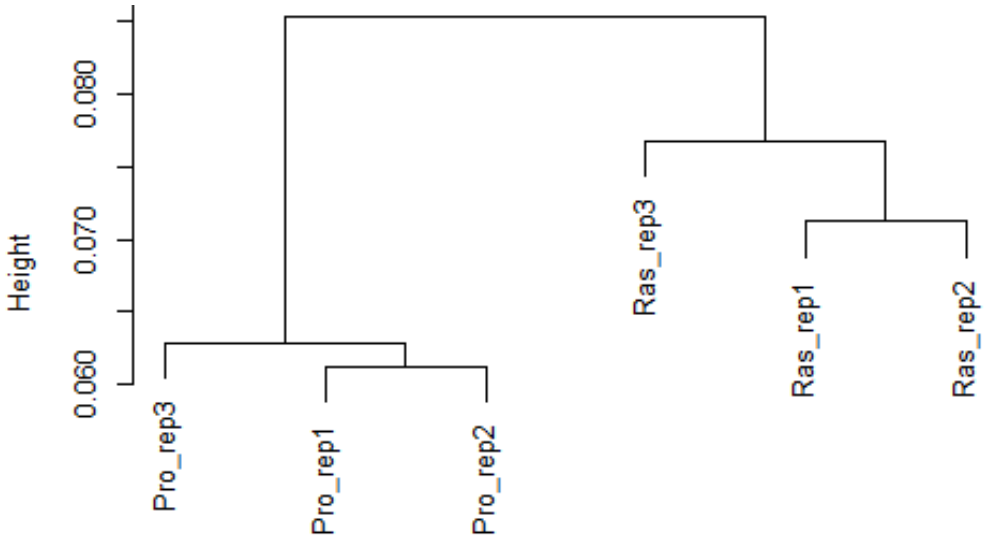
RPKMデータ

	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG00000000003	7.14	7.68	6.17	3.07	4.51	6.81
ENSG00000000005	0.00	0.00	0.00	0.05	0.00	0.00
ENSG00000000419	10.31	13.03	8.62	10.04	12.32	9.57
ENSG00000000457	1.92	2.29	2.02	2.98	3.19	1.85
ENSG00000000460	0.79	0.78	0.80	0.64	0.63	0.51
…						

*TCC*を用いた発現変動遺伝子同定

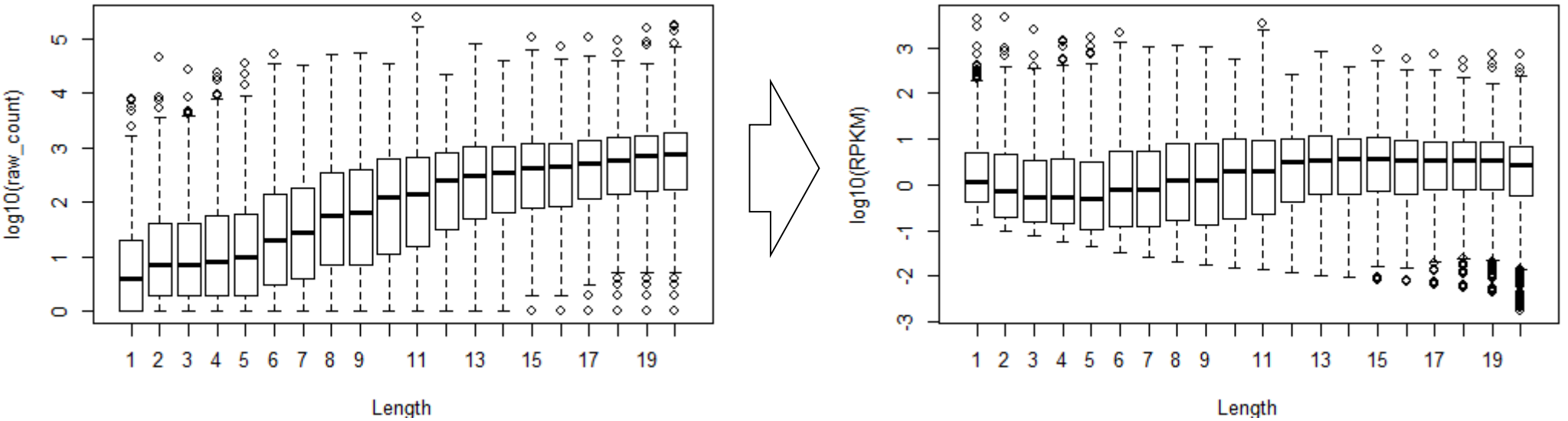


サンプル間クラスタリング



RPKMのデータ

- 配列長由来の偏りを補正
  - 現象：配列長が長いほどカウント数が増加傾向(Mortazavi et al., 2008)
  - 対策：Reads per kilobase (RPK)。配列長で割って1000を掛ける
- サンプル間での総リード数の違いを補正
  - サンプルごとにマップされた総リード数(総カウント数)が異なる
  - 対策：Reads per million (RPM)。総リード数で割って100万を掛ける



RP(K)M補正によって、length biasが軽減。転写物間の発現レベルの大小関係を知りたい場合に利用

発現変動解析用Rパッケージ

- *DEGSeq* (Wang *et al*., *Bioinformatics*, **26**: 136-138, 2
- *edgeR* (Robinson *et al*., *Bioinformatics*, **26**: 139-140,
- *GPseq* (Srivastava and Chen, *Nucleic Acids Res*., **38**
- *baySeq* (Hardcastle and Kelly, *BMC Bioinformatics*, **1**
- *DESeq* (Anders and Huber, *Genome Biol*., **11**: R106,
- *NBPSeq* (Di *et al*., *SAGMB*, **10**: article24, 2011)
- *TPSM* (Auer and Doerge, *SAGMB*, **10**: article26, 201
- *BBSeq* (Zhou *et al*., *Bioinformatics*, **27**: 2672-2678, 2
- *NOISeq* (Tarazona *et al*., *Genome Res*., **21**: 2213-22
- *PoissonSeq* (Li *et al*., *Biostatistics*, **13**: 523-538, 2012
- *SAMseq* (Li and Tibshirani, *Stat Methods Med Res*., 2
- *BitSeq* (Glaus *et al*., *Bioinformatics*, **28**: 1721-1728, 2
- *DEXSeq* (Anders et al., *Genome Res*., **22**: 2008-2017
- *ShrinkBayes* (Van DE Wiel *et al*., *Biostatistics*, **14**: 11
- *sSeq* (Yu *et al*., *Bioinformatics*, **29**: 1275-1282, 2013)
- *TCC* (Sun *et al*., *BMC Bioinformatics*, **14**: 219, 2013)
- …

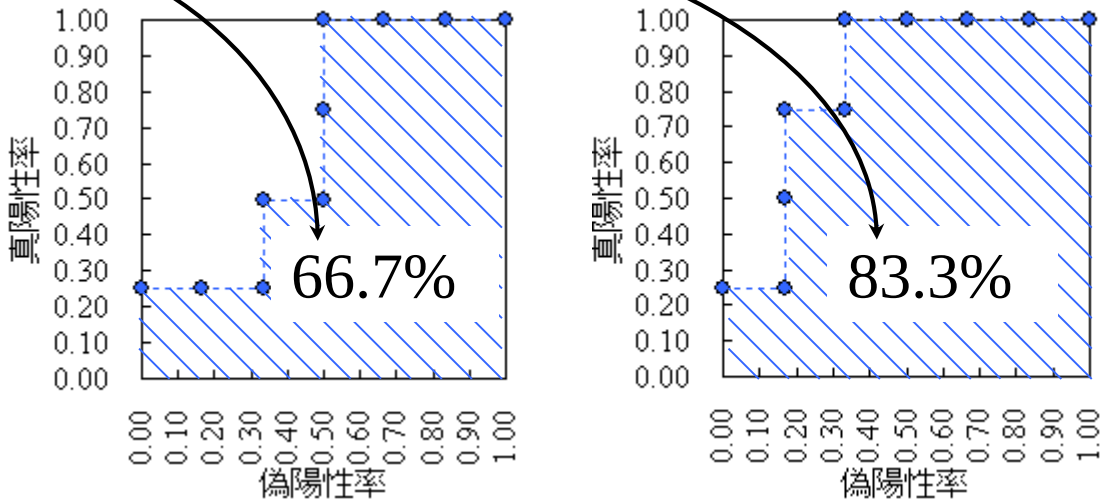
黒字のものたち（+α）の比較結果は…

- *DEGSeq* (Wang *et al*., *Bioinformatics*, **26**: 136-138, 2010)
- *edgeR* (Robinson *et al*., *Bioinformatics*, **26**: 139-140, 2010)
- *GPseq* (Srivastava and Chen, *Nucleic Acids Res*., **38**: e170, 2010)
- *baySeq* (Hardcastle and Kelly, *BMC Bioinformatics*, **11**: 422, 2010)
- *DESeq* (Anders and Huber, *Genome Biol*., **11**: R106, 2010)
- *NBPSeq* (Di *et al*., *SAGMB*, **10**: article24, 2011)
- *TPSM* (Auer and Doerge, *SAGMB*, **10**: article26, 2011)
- *BBSeq* (Zhou *et al*., *Bioinformatics*, **27**: 2672-2678, 2011)
- *NOISeq* (Tarazona *et al*., *Genome Res*., **21**: 2213-2223, 2011)
- *PoissonSeq* (Li *et al*., *Biostatistics*, **13**: 523-538, 2012)
- *SAMseq* (Li and Tibshirani, *Stat Methods Med Res*., 2011 Nov 28)
- *BitSeq* (Glaus *et al*., *Bioinformatics*, **28**: 1721-1728, 2012)
- *DEXSeq* (Anders et al., *Genome Res*., **22**: 2008-2017, 2012)
- *ShrinkBayes* (*ShrinkSeq*; Van DE Wiel *et al*., *Biostatistics*, **14**: 113-128, 2013)
- *sSeq* (Yu *et al*., *Bioinformatics*, **29**: 1275-1282, 2013)
- *TCC* (Sun *et al*., *BMC Bioinformatics*, **14**: 219, 2013)

よりよい方法とは？

- その方法を用いて発現変動の度合いでランキングしたときに、真の発現変動遺伝子（DEG）がより上位にランキングされる（感度・特異度高い）

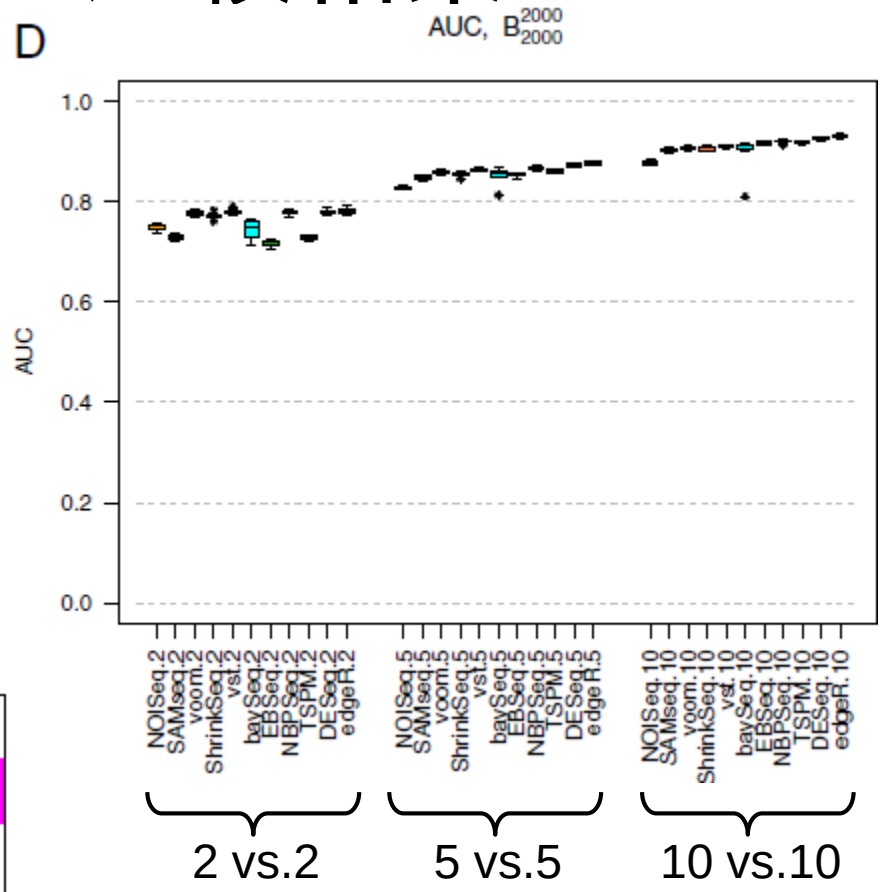
ランキング法				
rank	方法1		方法2	
1	gene8	真	gene8	真
2	gene5	偽	gene5	偽
3	gene4	偽	gene3	真
4	gene3	真	gene2	真
5	gene7	偽	gene7	偽
6	gene1	真	gene1	真
7	gene2	真	gene4	偽
8	gene9	偽	gene9	偽
9	gene10	偽	gene10	偽
10	gene6	偽	gene6	偽

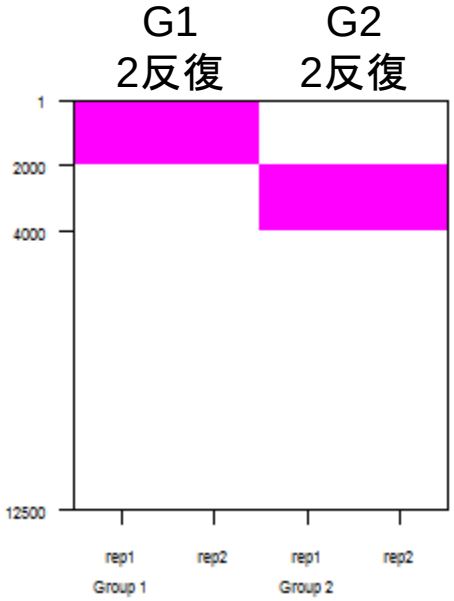


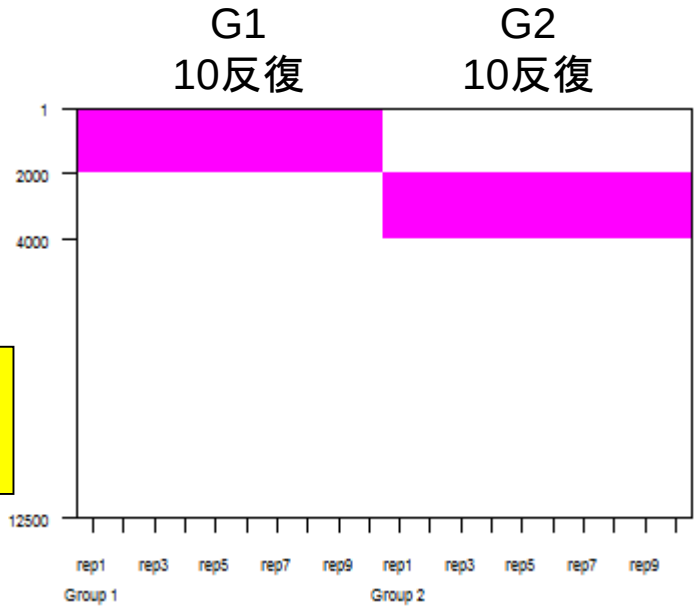
Area Under the ROC Curve (ROC曲線の下部面積：AUC)
バイオインフォマティクス分野でよく用いられる評価基準です

AUC値の比較結果

シミュレーション条件：G1 vs. G2
全遺伝子数：12500
発現変動遺伝子(DEG)数：4000
　G1で高発現：2000
　G2で高発現：2000
unbiased DE situation



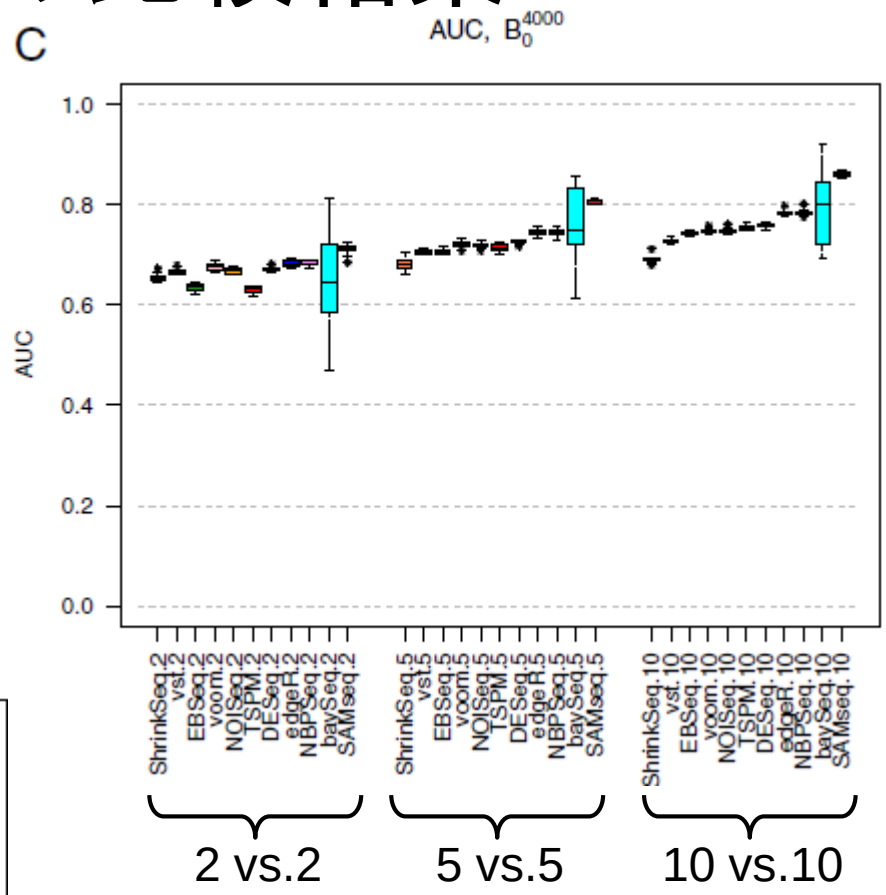


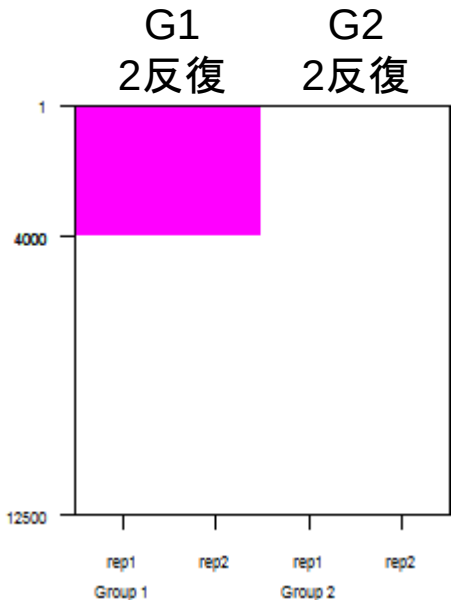


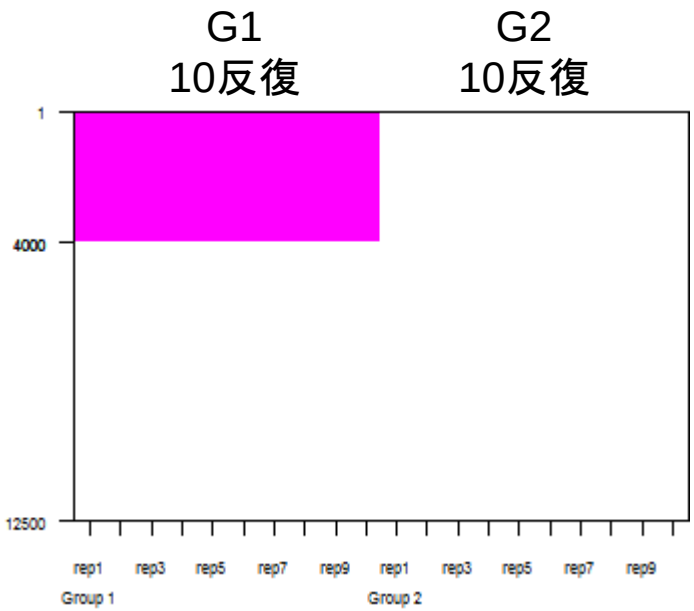
反復実験数を増やすほど精度は上がる
（これが言いたいわけではない…）

AUC値の比較結果

シミュレーション条件：G1 vs. G2
全遺伝子数：12500
発現変動遺伝子(DEG)数：4000
　G1で高発現：4000
　G2で高発現：0
biased DE situation





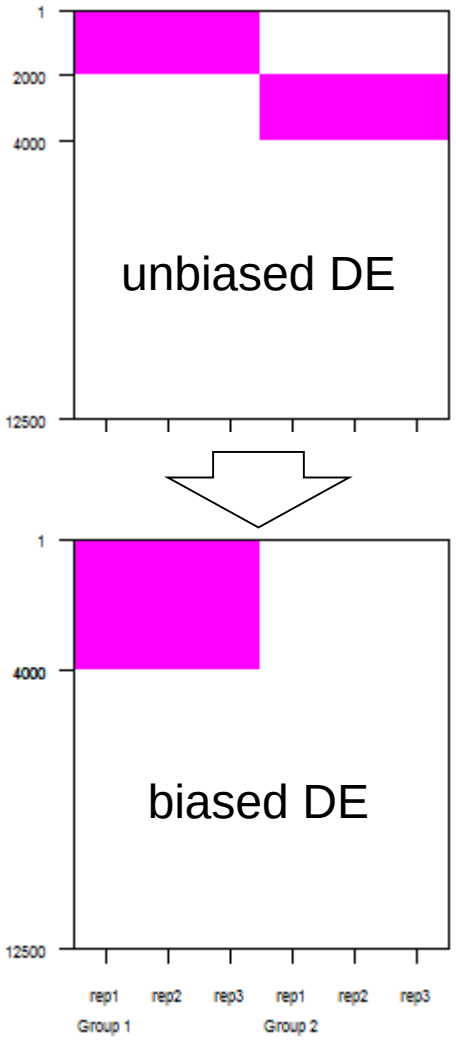


グループ（群）間でDEG数の組成に偏りがあると精度が大幅に低下する

理由：データ正規化法がDEG数の組成に偏りがないことを想定しているため

AUC値の比較結果

3反復 vs. 3反復



edgeR	SAMseq	*TCC*
90.84%	87.19%	90.83%

edgeR	SAMseq	*TCC*
82.95%	84.40%	89.92%

偏りのないデータのときは*edgeR*がよい
偏りのあるデータのときはSAMseqがよい
→ **偏りの有無に関係なく*TCC*がよい**

TCC（ver. 1.2.0）の主な機能

- DEG elimination strategy (DEGES)に基づくデータ**正規化**法を実装

  - 複製ありデータ用

	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3
gene_1	36	56	144	2	1	0
gene_2	84	152	124	52	37	28
gene_3	592	840	800	151	257	200
gene_4	0	8	4	1	1	3
…						

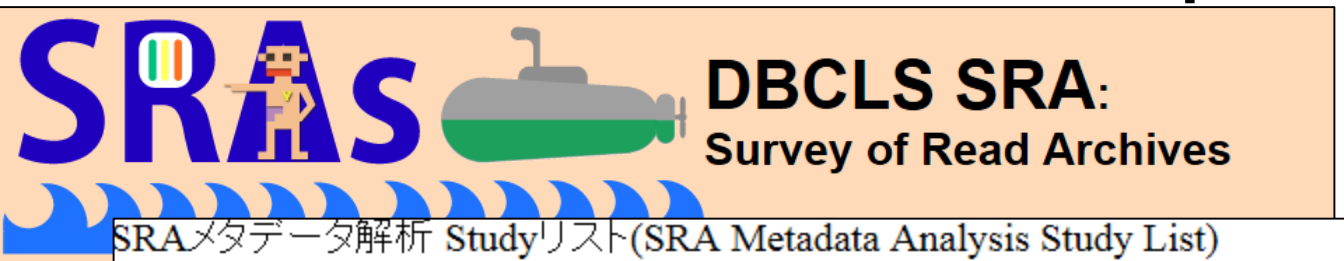
    - TbT正規化法（Kadota et al., 2012）: TMM-baySeq-TMMパイプライン
    - **iDEGES/edgeR正規化法**：TMM-(edgeR-TMM)$_n$パイプライン
  - 複製なしデータ用

	G1_rep1	G2_rep1
gene_1	36	2
gene_2	84	52
gene_3	592	151
gene_4	0	1
…		

    - **iDEGES/DESeq正規化法**：DESeq-(DESeq-DESeq)$_n$パイプライン
- 既存パッケージ中の**DEG検出**法を呼び出して利用可能
  - （正規化のところと同じく）*edgeR*, *baySeq*, *DESeq*パッケージ中の関数群を内部的に利用
- シミュレーションデータ作成機能
  - 二群（複製あり and/or なし）、三群、四群、、、多群
  - 発現変動の度合いを調整可能

*TCC*は（既存の手法を組み合わせることで）データ正規化部分の精度向上に貢献

カイコのsmall RNA-seq解析をRで…



SRAメタデータ解析 Studyリスト(SRA Metadata Analysis Study List)

→ back to SRAs top

RSS
検索条件設定(Search condition):クリア(Clear)
現在設定されている条件(Condition)
TYPE : Transcriptome Analysis
SCIENTIFIC_NAME : Bombyx mori

All (25828)	Transcriptome Analysis (3281)	Metagenomics (1896)	Epigenetics (1116)	Resequencing (20)	Gene Regulation Study (2)	Population Genomics (177)	RNASeq (82)	Cancer Genomics (44)	Forensic or Paleo-genomics (0)	Synthetic Genomics (8)	Whole Genome Sequencing (14978)	Other (4137)	検索結果(Result) (14)

pages: 1 records:14

No.	SRA#	Study#	TITLE	STUDY TYPE	Exps	Runs	UPDATE DATE
7	DRA000943	DRP000980	Hsp90 facilitates accurate loading of precursor piRNAs into PIW teins	Transcriptome Analysis	6	6	2013-03-06
13	SRA060712	SRP016842	GSE41841: Bombyx mori Argonaute2 associated small RNAs	Transcriptome Analysis	1	1	2012-10-26
12	SRA055759	SRP014173	GSE39203: A role for Fkbp6 and the chaperone machinery in piR A silencing				
5	DRA000527	DRP000552	The transcriptome in BmN4 cell line				
6	DRA000527	DRP000553	The massive transcriptional start site in BmN4 cell line				
11	SRA046010	SRP008285	Trancriptome analysis of the Interaction between Bombyx mori Nucl				
10	SRA040340	SRP007541	Alternative splicing and trans-splicing events revealed by analysis of				
14	SRA091413	SRP026212	GSE48168: The Small RNA Profile of Silkworm Hemolymph Using Deep Sequencing	Transcriptome Analysis	2	2	2011-07-07
3	DRA000374	DRP000378	De novo production of PIWI-interacting RNAs against a protein coding gene	Transcriptome Analysis	3	3	2011-04-05
2	DRA000317	DRP000318	Large-scale profiling of piRNAs deriving from silkworm developing embryos	Transcriptome Analysis	5	5	2011-01-03
4	DRA000275	DRP000276	Profiling of PIWI-interacting RNAs from the W chromosome-linked mutant that shows sex differentiation deficiency in females	Transcriptome Analysis	3	3	2010-09-15
1	DRA000173	DRP000173	Identification of PIWI-interacting RNAs deriving from the silkworm W chromosome	Transcriptome Analysis	5	5	2010-05-26
9	SRA012426	SRP002273	Single-Base Resolution Methylomes of Silkworms and Functional Importance of Gene Methylation in Insects Revealed by Ultra-High-Throughput Sequencing	Transcriptome Analysis	2	2	2010-04-09
8	SRA011055	SRP001890	MicroRNAs of Bombyx mori identified by Solexa sequencing	Transcriptome Analysis	3	3	2010-02-18

このsRNAデータのダウンロードからカイコゲノムへのマッピングまでの一連の手順を「**(Rで)塩基配列解析**」に掲載してあります。

(Rで)塩基配列解析(主にNGSやRNA-seq解析)

(last modified 2013/10/28, since 2010)

What's new?

- コード内のオプション名のところ(param...周辺)の統一化を行っています。(2013/10/21) NEW
- 一連の解析パイプライン(RNA-seqデータ取得 > マッピング > カウントデータやRPKMデータ取得 > サンプル間クラスタリングや発現変動解析およびM-A plot描画まで
- 発現変動解析用Rパッケージ TCC (
利用したい方は、R (ver. 3.0.2)をイン
ールと起動を参考にして、通常のイ
- BioMart周辺の情報をアップデートし
- 3群間比較解析(single-factorのみ)を
- 2013年10月30日13:30-15:00に開催
そして創薬」のフォーカストセッション
話す予定です。(2013/10/08) NEW
- (かなり先の話ですが...)平成26年3
行います。情報はかなりアップデー
- どのブラウザからでもエラーなく見
- 「(Rで)マイクロアレイデータ解析」も
- 2013年7月29日まで公開していた以
す(110MB程度)。(2013/07/30)
- 2013年6月6日に開催されたNAIST
TCC(ver. 1.1.99)などのインストールベルで作成した(しつつある)ものです。

- はじめに (last modified 2013/10/26
- Rのインストールと起動(last modifi
- サンプルデータ(last modified 2013
- イントロ | 一般 | ランダムに行を抽出
- イントロ | 一般 | 任意の文字列を行
- イントロ | 一般 | 任意のキーワードを
- イントロ | 一般 | ランダムな塩基配列
- イントロ | 一般 | 任意の長さの可能
- イントロ | 一般 | 任意の位置の塩基
- イントロ | 一般 | 指定した範囲の配
- イントロ | 一般 | 翻訳配列(translate)

- 作図 | について(last modified 2012/09/10)
- 作図 | M-A plot(基本編)(last modified 2012/10/01)
- 作図 | M-A plot(ggplot2編)(last modified 2013/07/30)
- 作図 | ROC曲線(last modified 2012/10/01)
- 作図 | SplicingGraphs(last modified 2013/08/07)
- パイプライン | について(last modified 2013/10/17) NEW
- パイプライン | ゲノム | 発現変動 | 2群間 | 対応なし | 複製あり | SRP017142(Neyret-Kahn_2013)(last modified 2013/10/23) NEW
- パイプライン | ゲノム | small RNA | SRP016842(Nie_2013)(last modified 2013/10/26) NEW
- リンク集(last modified 2012/03/29)

はじめに NEW

このページは、次世代シーケンサー(N
ための一連の手続きをまとめているも
Maintainerは門田幸二(東京大学大学
の出身研究室(東京大学・大学院農学
validation、美しいコーディング、TCCパ
アグリバイオの分野は、ヒトやマウスと
扱います。例えば、主に次世代シーケ
を行い、発現変動領域を同定する」と
ラムを利用するしかありませんが、201
ロアレイ解析を行うためにRに慣れたW
派ではありませんでしたが、50-100人
験上、「Rでコピペ」を基本とする以外
生がつまづきます。「スペース」、「特殊
な局面に遭遇し、学生の側も手を動か
像の理解」もままならずストレス、とい
心穏やかにハンズオン講義を行うため
小限の労力で解析したい人」にとって
ここでは、アグリバイオインフォマティク
化の遺伝子基盤解明」の研究を遂行す
など一切の保証はできませんので予め
ようなものです。リンクは自由にしてい
一切必要ありません。個別の項目への

パイプライン | ゲノム | small RNA | SRP016842(Nie_2013) NEW

Nie et al., BMC Genomics, 2013のカイコ (Bombyx mori) small RNA-seqデータが GSE41841に登録されています。(そしてリンク先のGSM1025527からも様々な情報を得ることができます。) ここでは、SRAdbパッケージを用いたそのFASTQ形式ファイルのダウンロードから、QuasRパッケージを用いたマッピングまでを行う一連の手順を示します。
basic alignerの一つであるBowtieをQuasRの内部で用いています。
ここでは「デスクトップ」上に「SRP016842」というフォルダを作成しておき、そこで作業を行うことにします。

Step1. RNA-seqデータのgzip圧縮済みのFASTQファイルをダウンロード:

論文中の記述からGSE41841を頼りに、SRP016842にたどり着いています。したがって、ここで指定するのは"SRP016842"となります。
以下を実行して得られるsmall RNA-seqファイルは一つ(SRR609266.fastq.gz)で、ファイルサイズは400Mb弱、11928428リードであることがわかります。
イントロ | NGS | 配列取得 | FASTQ or SRALite | SRAdb(Zhu_2013)の記述内容と基本的に同じです。

```
param <- "SRP016842"                    #取得したいSRA IDを指定

#必要なパッケージをロード
library(SRAdb)                          #パッケージの読み込み

#前処理
#sqlfile <- "SRAmetadb.sqlite"          #最新でなくてもよく、手元に予めダウンロードしてある"SRAmetadb.sql
sqlfile <- getSRAdbFile()               #最新のSRAmetadb SQLiteファイルをダウンロードして解凍(圧縮状態で3
sra_con <- dbConnect(SQLite(), sqlfile)#おまじない

#前処理(実験デザインの全体像を表示)
hoge <- sraConvert(param, sra_con=sra_con)#paramで指定したSRA IDに付随するstudy (SRP...), sample(SRS..
hoge                                    #hogeの中身を表示

#前処理(FASTQファイルサイズを表示)
k <- getFASTQinfo(hoge$run)             #「hoge$run」で指定したSRRから始まるIDのFASTQファイルサイズ情報な
k                                       #kの中身を表示
hoge2 <- cbind(k$library.name,          #ライブラリ名と、
               k$run.read.count,        #総リード数と、
               k$file.name,             #ファイル名と、
```

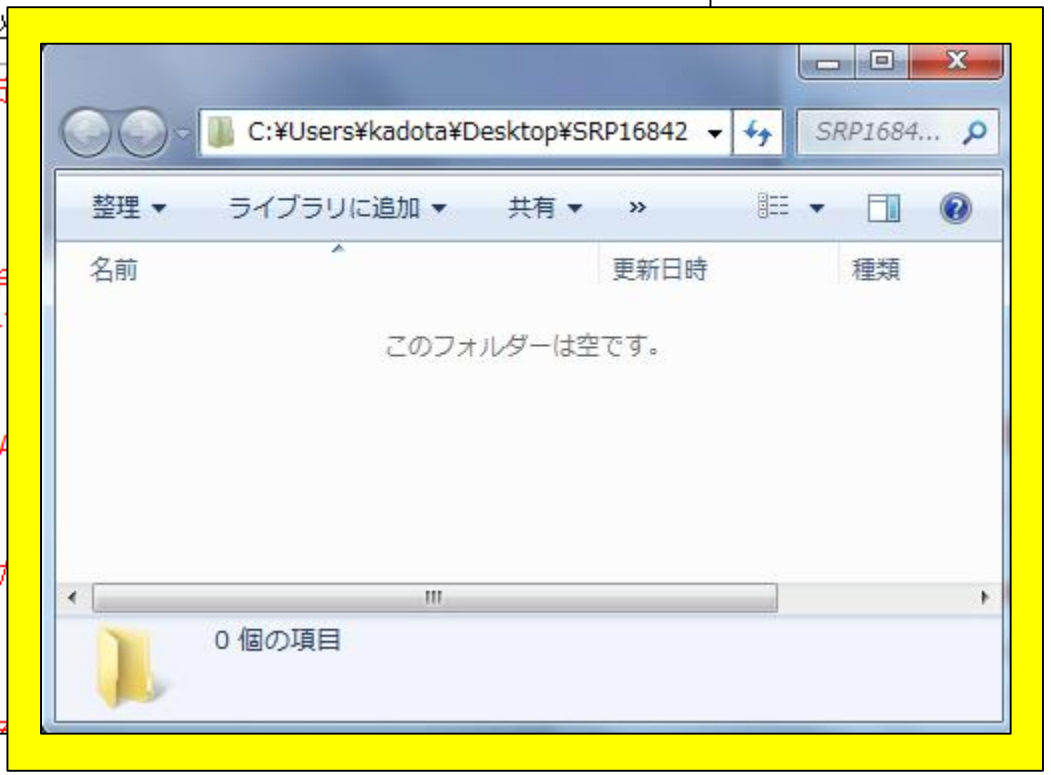
パイプライン | ゲノム | small RNA | SRP016842(Nie_2013) NEW

Nie et al., BMC Genomics, 2013のカイコ (Bombyx mori) small RNA-seqデータが GSE41841に登録されています。(そしてリンク先のGSM1025527からも様々な情報を得ることができます。) ここでは、SRAdbパッケージを用いたそのFASTQ形式ファイルのダウンロードから、QuasRパッケージを用いたマッピングまでを行う一連の手順を示します。
basic alignerの一つであるBowtieをQuasRの内部で用いています。
ここでは「デスクトップ」上に「SRP016842」というフォルダを作成しておき、そこで作業を行うことにします。

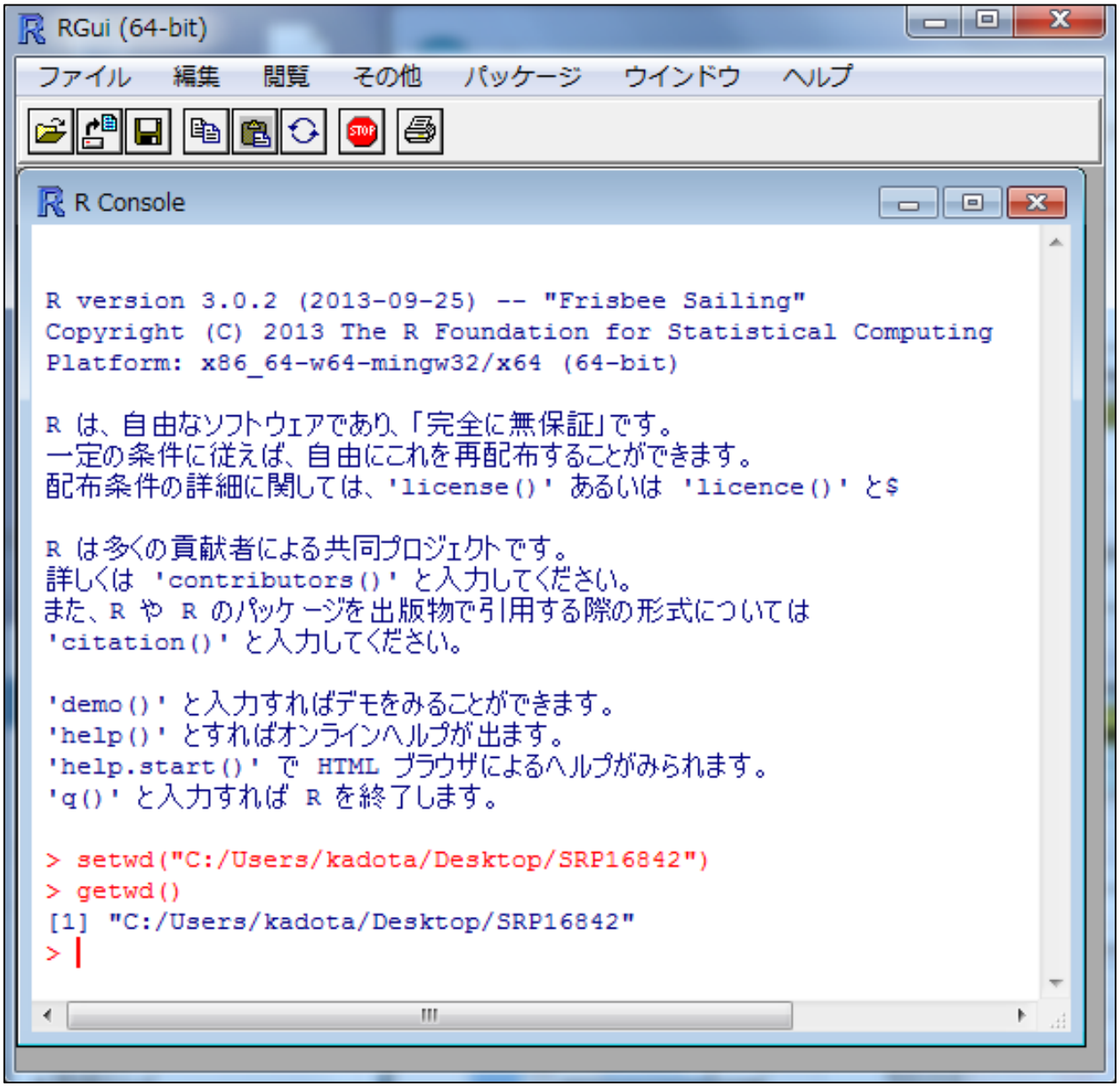
Step1. RNA-seqデータのgzip圧縮済みのFASTQファイルをダウンロード:

論文中の記述からGSE41841を頼りに、SRP016842にたどり着いています。したがって、ここで指定するのは"SRP016842"となります。
以下を実行して得られるsmall RNA-seqファイルは一つ(SRR609266.fastq.gz)で、ファイルサイズは400Mb弱、11928428リードであることがわかります。
イントロ | NGS | 配列取得 | FASTQ or SRALite | SRAdb(Zhu_2013)の記述内容と

```
param <- "SRP016842"                          #取得したいSRA IDを指定

#必要なパッケージをロード
library(SRAdb)                                #パッケージの読み込み

#前処理
#sqlfile <- "SRAmetadb.sqlite"                #最新でなくてもよく、手
sqlfile <- getSRAdbFile()                     #最新のSRAmetadb SQLi
sra_con <- dbConnect(SQLite(), sqlfile)#おまじない

#前処理(実験デザインの全体像を表示)
hoge <- sraConvert(param, sra_con=sra_con)#paramで指定したSRA
hoge                                          #hogeの中身を表示

#前処理(FASTQファイルサイズを表示)
k <- getFASTQinfo(hoge$run)                   #「hoge$run」で指定した
k                                             #kの中身を表示
hoge2 <- cbind(k$library.name,                #ライブラリ名と、
               k$run.read.count,              #総リード数と、
               k$file.name,                   #ファイル名と、
```



Rを起動して作業ディレクトリの変更



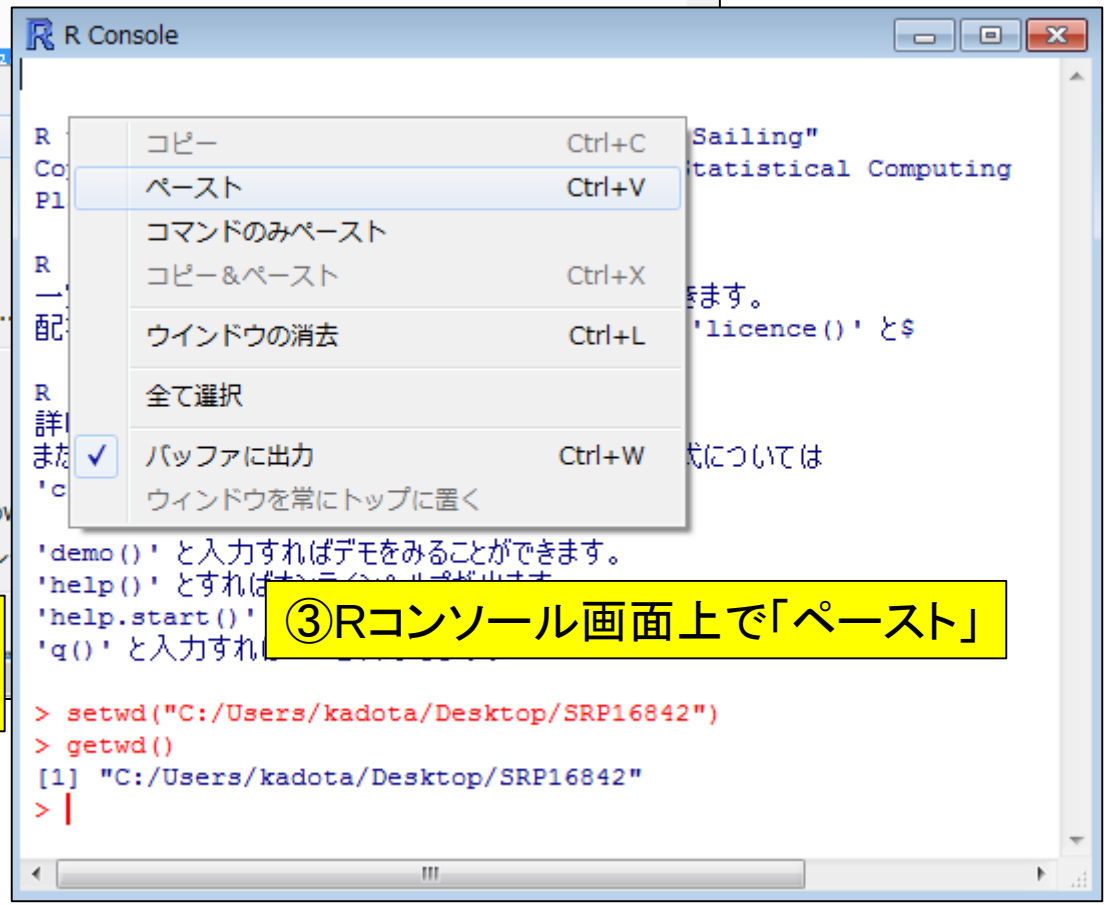
Step1：sRNA-seqファイルのダウンロード

Step1. RNA-seqデータのgzip圧縮済みのFASTQファイルをダウンロード：

論文中の記述からGSE41841を頼りに、SRP016842にたどり着いています。したがって、ここで指定するのは"SRP016842"となります。以下を実行して得られるsmall RNA-seqファイルは一つ(SRR609266.fastq.gz)で、ファイルサイズは400Mb弱、11928428リードであることがわかります。

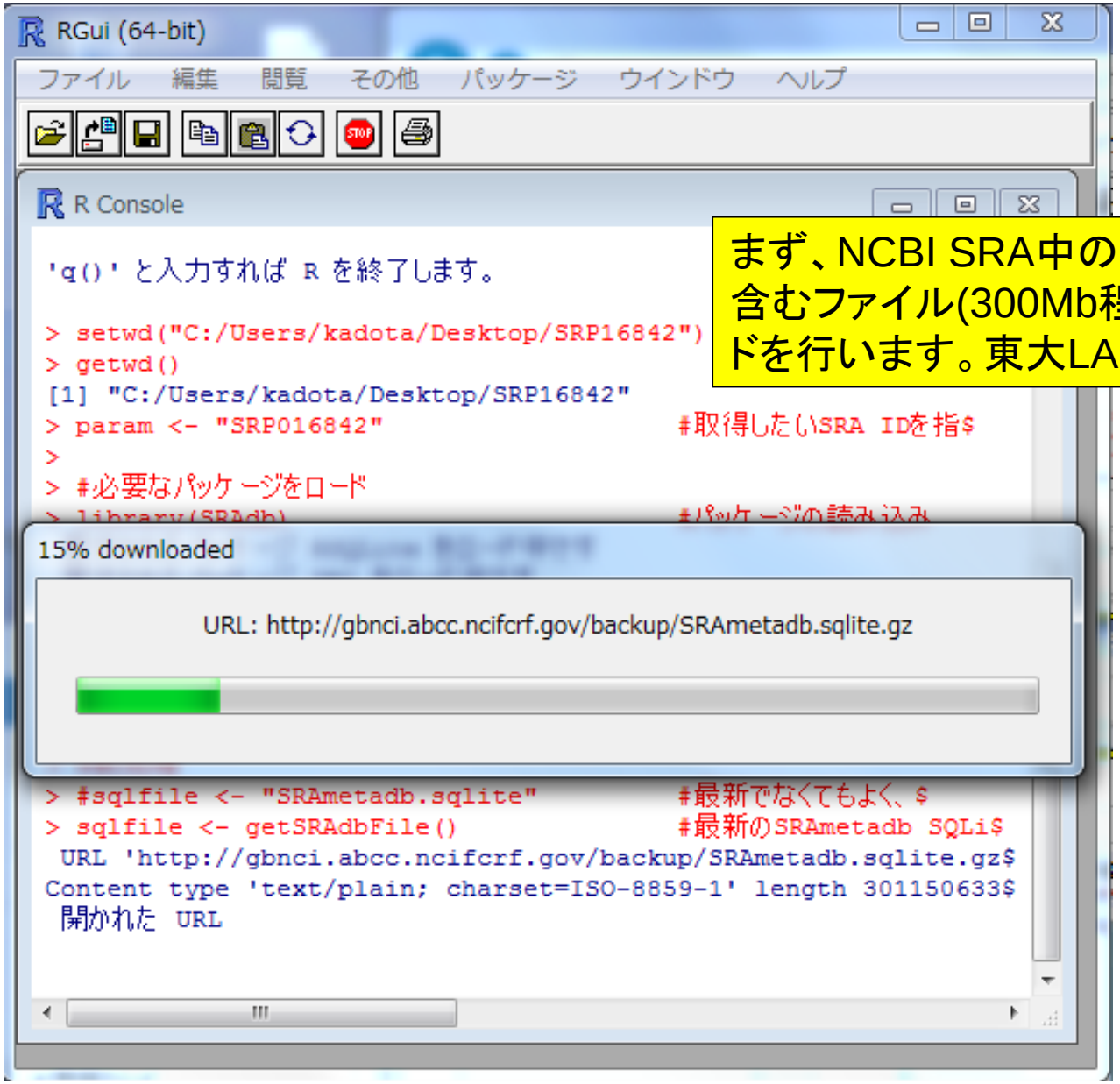
イントロ | NGS | 配列取得 | FASTQ or SRALite | SRAdb(Zhu 2013)の記述内容と基本的に同じです。

```
param <- "SRP016842"                    #取得したいSRA IDを指定

#必要なパッケージをロード
library(SRAdb)                          #パッケージの読み込み

#前処理
#sqlfile <- "SRAmetadb.sqlite"          #
sqlfile <- getSRAdbFile()               #
sra_con <- dbConnect(SQLite(), sqlfile)#

#前処理(実験デザインの全体像を表示)
hoge <- sraConvert(param, sra_con=sra_co
hoge                                    #

#前処理(FASTQファイルサイズを表示)
k <- getFASTQinfo(hoge$run)             #
k                                       #
hoge2 <- cbind(k$library.name,          #
               k$run.read.count,        #
               k$file.name,             #
```



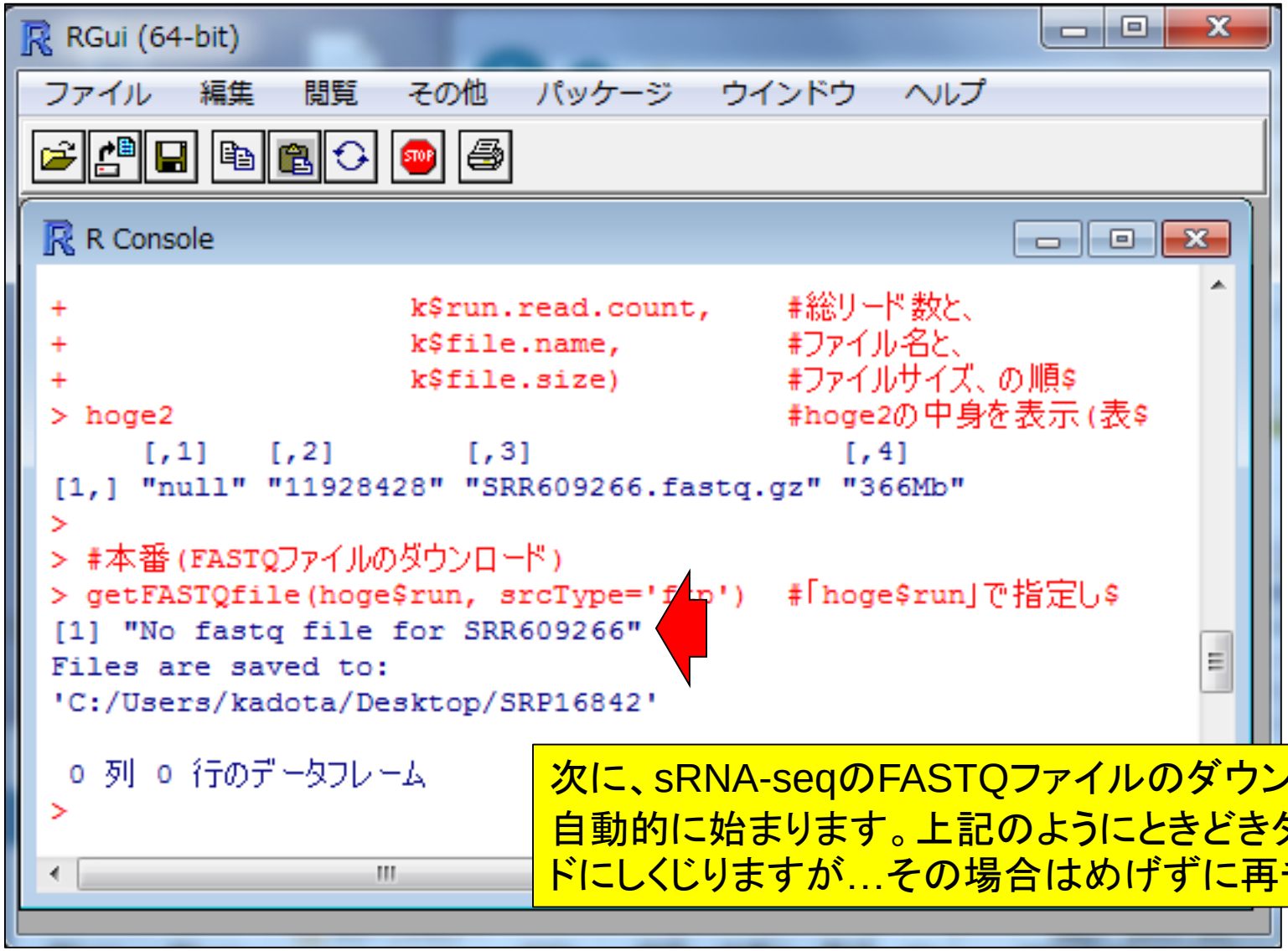
①CTRLキーを押しながら左クリックで全選択
②右クリックで「コピー」

Step1：sRNA-seqファイルのダウンロード

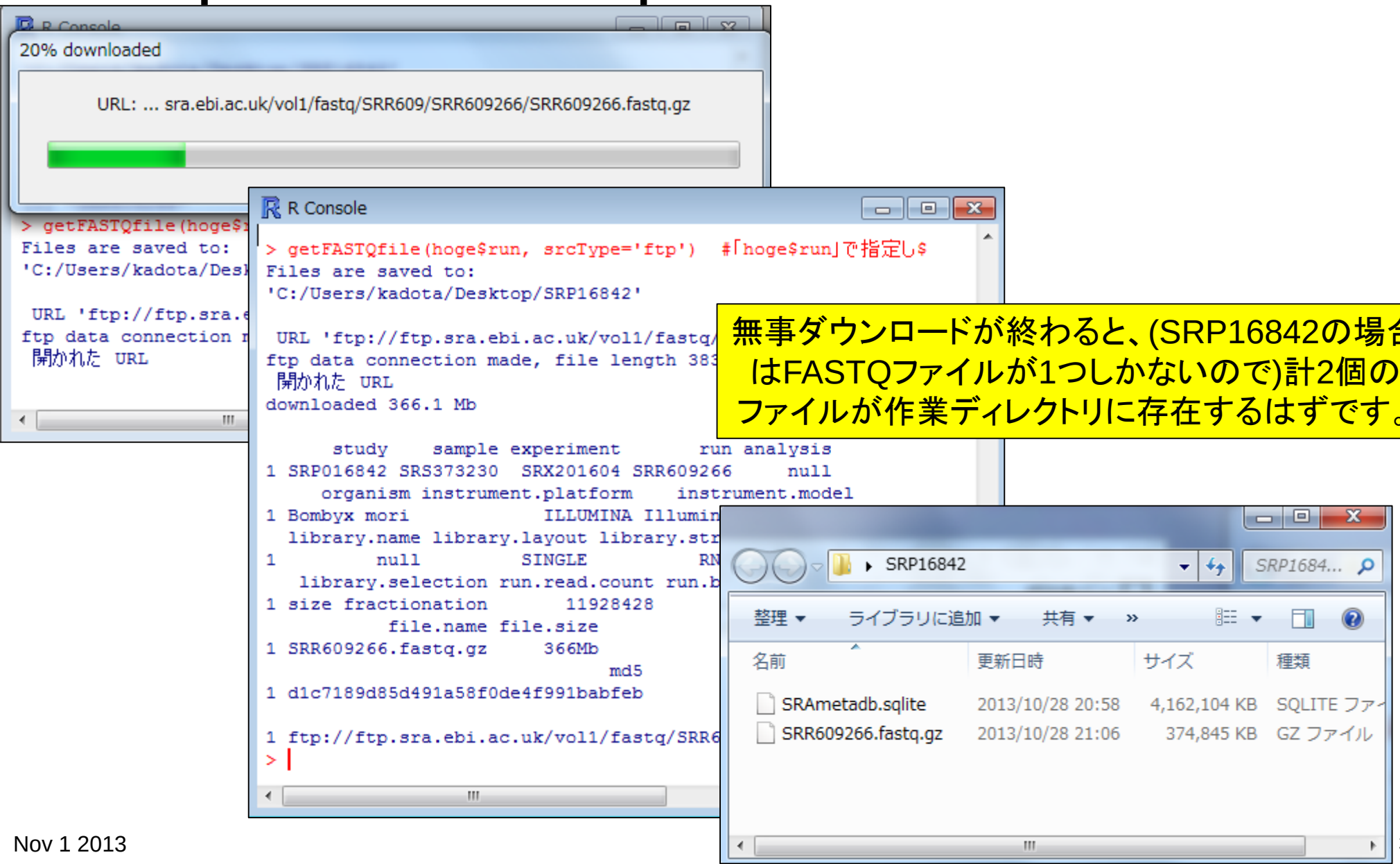


まず、NCBI SRA中のメタデータ情報を含むファイル(300Mb程度)のダウンロードを行います。東大LANで10分程度。

Step1：sRNA-seqファイルのダウンロード



Step1：sRNA-seqファイルのダウンロード



無事ダウンロードが終わると、(SRP16842の場合はFASTQファイルが1つしかないので)計2個のファイルが作業ディレクトリに存在するはずです。

Step2：リファレンス配列のダウンロード

Step2. カイコゲノム配列をダウンロード：

カイコゲノム配列はBSgenomeパッケージとして提供されていないため、自力で入手する必要があります。手順としては、農業生物資源研究所(NIAS)が提供しているカイコゲノム配列のウェブページからIntegrated sequences (integretedseq.txt.gz) をダウンロードし、解凍します。解凍後のファイル名は"integretedseq.txt"となりますが、拡張子を".txt"から".fa"に変更して、"integretedseq.fa"としたものを作業ディレクトリ(「デスクトップ」上の「SRP016842」フォルダ)上にコピーしておきます。

Step3：前処理(アダプター配列除去など)

Step3. small RNA-seqデータの前処理:

原著論文(Nie et al., 2013)中では、アダプター配列やクオリティの低いリードを除去したのち、ゲノムにマッピングしたと書いてあります。アダプター配列情報はどこにも書かれていませんでしたが、Table S2中のアダプター配列除去後の最も短いリードが18 nt (例:"GCAGTCGTGGCCGAGCGG")であり、「この18 nt」と「この配列を含む生リード配列の差分」がアダプター配列ということになります。詳細な情報は書かれていませんでしたが、おそらくアダプター配列は"TGGAATTCTCGGGTGCCAAGGAACTCCAGTC..."という感じだろうと推測できます。

ここでは、ダウンロードした"SRR609266.fastq.gz"ファイルを入力として、1塩基ミスマッチまで許容して(推定)アダプター配列除去を行ったのち、"ACGT"のみからなる配列(許容するN数が0)で、配列長が18nt以上のものをフィルタリングして出力しています。

```
in_f <- "SRR609266.fastq.gz"            #入力ファイル名を指定してin_fに格納(RNA-seqファイル)
out_f <- "SRR609266_p.fastq.gz"         #出力ファイル名を指定してout_fに格納
param_adapter <- "TGGAATTCTCGGGTGCCAAGGAACTCCAGTC"#アダプター配列を指定
param_mismatch <- 1                     #許容するミスマッチ数を指定
param_nBases <- 0                       #許容するNの数を指定
param_minLength <- 18                   #アダプター配列除去後の許容する最低配列長を指定

#必要なパッケージをロード
library(QuasR)                          #パッケージの読み込み

#本番(前処理)
res <- preprocessReads(filename=in_f,   #前処理を実行
             outputFilename=out_f,      #前処理を実行
             Rpattern=param_adapter,    #前処理を実行
             max.Rmismatch=rep(param_mismatch, nchar(param_adapter)),#前処理を実行
             nBases=param_nBases,       #前処理を実行
             minLength=param_minLength)#前処理を実行
res                                     #確認してるだけです
```

前処理実行後のresオブジェクトを眺めると、入力ファイルのリード数が11928428であり、アダプター配列除去後に18nt未満の長さになってしまったためにフィルタリングされたリード数が157229、"N"を1つ以上含むためにフィルタリングされたリード数が21422あったことがわかります。ここでは配列長分布は得ておりませんが、出力ファイルを解凍して配列長分布を調べると原著論文中のTable S1と似た結果になっていることから、ここでの処理が妥当であることがわかります。

Step3：前処理(アダプター配列除去など)

Step3. small RNA-seqデータの前処理:

原著論文(Nie et al., 2013)中では、アダプター配列やクオリティの低いリードを除
す。アダプター配列情報はどこにも書かれていませんでしたが、Table S2中のア
(例:"GCAGTCGTGGCCGAGCGG")であり、「この18 nt」と「この配列を含む生
す。詳細な情報は書かれていませんでしたが、おそらくアダプター配列は"TGC
感じだろうと推測できます。
ここでは、ダウンロードした"SRR609266.fastq.gz"ファイルを入力として、1塩基
ったのち、"ACGT"のみからなる配列(許容するN数が0)で、配列長が18nt以上

```
in_f <- "SRR609266.fastq.gz"              #入力ファイル名を指定
out_f <- "SRR609266_p.fastq.gz"           #出力ファイル名を指定
param_adapter <- "TGGAATTCTCGGGTGCCAAGGAACTCCAGTC"#アダプタ
param_mismatch <- 1                       #許容するミスマッチ数
param_nBases <- 0                         #許容するNの数を指定
param_minLength <- 18                     #アダプター配列除去後

#必要なパッケージをロード
library(QuasR)                            #パッケージの読み込み

#本番(前処理)
res <- preprocessReads(filename=in_f,     #前処理を実行
             outputFilename=out_f,        #前処理を実行
             Rpattern=param_adapter,      #前処理を実行
             max.Rmismatch=rep(param_mismatch, nchar(param_
             nBases=param_nBases,         #前処理を実行
             minLength=param_minLength)#前処理を実行
res                                       #確認してるだけです
```

R Console

```
    tapply, union, unique, unlist

 要求されたパッケージ IRanges をロード中です
 要求されたパッケージ XVector をロード中です
 要求されたパッケージ Rbowtie をロード中です
>
> #本番(前処理)
> res <- preprocessReads(filename=in_f,   #前処理を実行
+              outputFilename=out_f,      #前処理を実行
+              Rpattern=param_adapter,    #前処理を実行
+              max.Rmismatch=rep(param_mismatch, nchar(param_a$
+              nBases=param_nBases,       #前処理を実行
+              minLength=param_minLength)#前処理を実行
  filtering SRR609266.fastq.gz
> res                                     #確認してるだけです
                   SRR609266.fastq.gz
totalSequences               11928428
matchTo5pAdapter                   0
matchTo3pAdapter             11928428
tooShort                       157229
tooManyN                        21422
lowComplexity                      0
totalPassed                  11749931
>
```

5分ほどで終わります

前処理実行後のresオブジェクトを眺めると、入力ファイルのリード数が11928428であり、アダプター配列除去後に18nt未満の長さになってしまったためにフィルタリングされたリード数が157229、"N"を1つ以上含むためにフィルタリングされたリード数が21422あったことがわかります。ここでは配列長分布は得ておりませんが、出力ファイルを解凍して配列長分布を調べると原著論文中のTable S1と似た結果になっていることから、ここでの処理が妥当であることがわかります。

アダプター配列除去後の配列長分布

- イントロ | ファイル形式の変換 | qseq --> Illumina FASTQ (last modified 2013/06/17)
- イントロ | ファイル形式の変換 | qseq --> Sanger FASTQ (last modified 2013/08/19)
- 前処理 | クオリティチェック | について (last modified 2013/06/17)
- 前処理 | クオリティチェック | qrqc (last modified 2013/06/18)
- 前処理 | クオリティチェック | PHREDスコアに変換 (last modified 2013/06/18)
- 前処理 | クオリティチェック | 配列長分布を調べる (last modified 2013/06/18)
- 前処理 | フィルタリング | PHREDスコアが低い塩基をNに置換 (last modified
- 前処理 | フィルタリング | PHREDスコアが低い配列(リード)を除去 (last modi
- 前処理 | フィルタリング | ACGTのみからなる配列を抽出 (last modified 2013
- 前処理 | フィルタリング | ACGT以外のcharacter"-"をNに変換 (last modified
- 前処理 | フィルタリング | ACGT以外の文字数が閾値以下の配列を抽出 (las
- 前処理 | フィルタリング | 重複のない配列セットを作成 (last modified 2013/06
- 前処理 | フィルタリング | 指定した長さ以上の配列を抽出 (last modified 2013
- 前処理 | フィルタリング | 指定した長さの範囲の配列を抽出 (last modified 20
- 前処理 | フィルタリング | 任意のIDを含む配列を抽出 (last modified 2013/06
- 前処理 | フィルタリング | Illuminaのpass filtering (last modified 2013/06/19)
- 前処理 | フィルタリング | GFF/GTF形式ファイル (last modified 2013/10/10) N
- 前処理 | フィルタリング | 組合せ | ACGTのみ & 指定した長さの範囲の配列
- 前処理 | トリミング | ポリA配列除去 (last modified 2013/06/18)
- 前処理 | トリミング | アダプター配列除去(基礎) (last modified 2013/08/01)

前処理 | クオリティチェック | 配列長分布を調べる

FASTAまたはFASTQ形式ファイルを読み込んで配列長分布を得るやり方を示します。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動

1. FASTA形式ファイル(sample2.fasta)の場合:

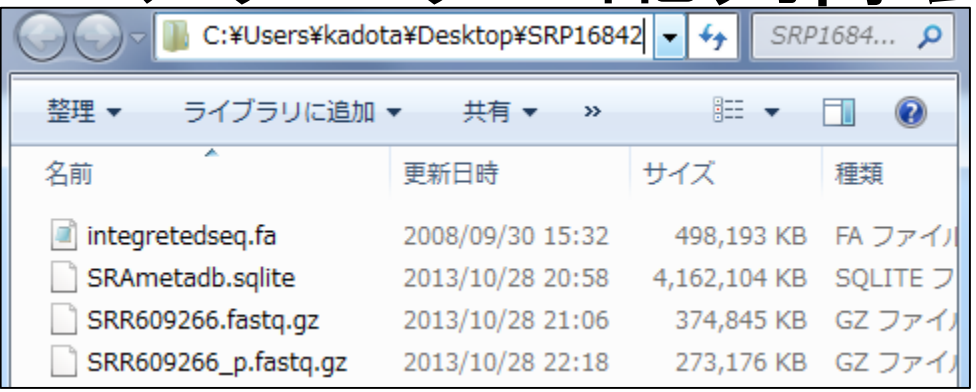
```
in_f <- "sample2.fasta"                    #入力ファイル名を指定してi
out_f <- "hoge1.txt"                       #出力ファイル名を指定してo

#必要なパッケージをロード
library(Biostrings)                        #パッケージの読み込み

#入力ファイルの読み込み
fasta <- readDNAStringSet(in_f, format="fasta")#in_fで指定したフ
fasta                                      #確認してるだけです

#本番
out <- table(width(fasta))                 #長さごとの出現頻度情報を得
out                                        #確認してるだけです

#ファイルに保存
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=
```

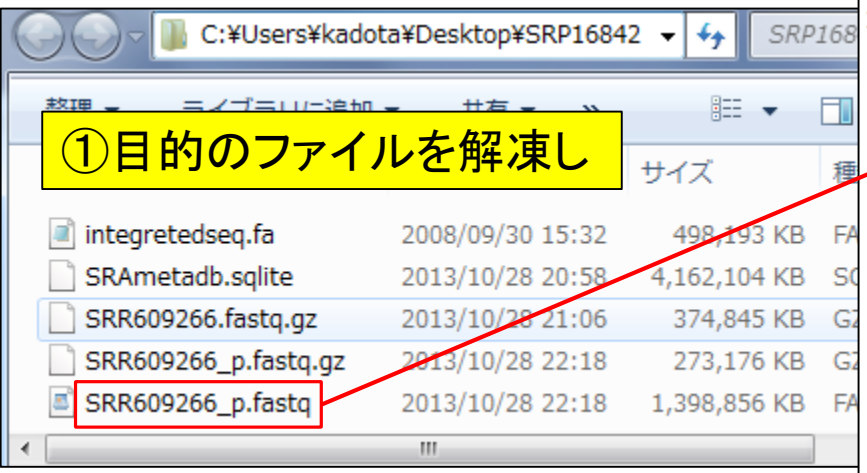
手元のファイルはFASTQ圧縮ファイルなので…

2. FASTQ形式ファイル(sample2.fastq)の場合:

```
in_f <- "sample2.fastq"                    #入力ファイル名を指定してi
out_f <- "hoge2.txt"                       #出力ファイル名を指定してo

#必要なパッケージをロード
library(ShortRead)                         #パッケージの読み込み
```

アダプター配列除去後の配列長分布





2. FASTQ形式ファイル(sample2.fastq)の場合:

```
in_f <- "sample2.fastq"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge2.txt"                       #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(ShortRead)

#入力ファイルの読み込み
fastq <- readFastq(in_f)
sread(fastq)

#本番
out <- table(width(fastq))                 #長さごとの出現頻度情報を得た結果をou
out                                        #確認してるだけです

#ファイルに保存
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=F)#outの中
```

②「FASTQファイルの場合」をテンプレートにして、入力ファイルのところを書き換えてコピペ

アダプター配列除去後の

R Console

```
> sread(fastq)                              #配列情報を表示
  A DNAStringSet instance of length 11749931
            width seq
       [1]     27 TATGTCTAAGGAGAATTCAAAAAAGAG
       [2]     33 TCTTCGGTAGTATAGTGGTCAGTATCCCCGCCT
       [3]     33 TCTTCGGTAGTATAGTGGTCAGTATCCCCGCCT
       [4]     33 TCTTCGGTAGTATAGTGGTCAGTATCCCCGCCT
       [5]     22 CCTTGCAAACTGTAAACTCATT
       ...    ... ...
[11749927]     32 GCCGTGATCGTCTAGTGGTTAGGACCCTACGT
[11749928]     48 TTCTTCACAATTCGGCACAATG...CTCGGGTGCCCAGGAACTACA
[11749929]     33 AAGGGAGATATGGTTCGGAACGCGAAGAGCACC
[11749930]     33 TGCCGTGATCGTCTAGTGGTTAGGACCCTACGT
[11749931]     33 TCTTCGGTAGTATAGTGGTCAGTATCCCCGCCT
>
> #本番
> out <- table(width(fastq))                #長さごとの出現頻度情$
> out                                       #確認してるだけです

     18      19      20      21      22      23      24
  55040  104374  515188  264863  296527  163167  137087
     25      26      27      28      29      30      31
 172538  195653  337547  335273  131190   58490   68368
     32      33      34      35      36      37      38
1534331 6492940  135961   72352   73566   99069   63186
     39      40      41      42      43      44      45
  71399   66389   60610   36650   27401   30700   15275
     46      47      48
  16579   26789   91429
>
> #ファイルに保存
> write.table(out, out_f, sep="\t", append=F, quote=F, row.nam$
>
```

hoge2.txt

Length	Total
18	55040
19	104374
20	515188
21	264863
22	296527
23	163167
24	137087
25	172538
26	195653
27	337547
28	335273
29	131190
30	58490
31	68368
32	1534331
33	6492940
34	135961
35	72352
36	73566
37	99069
38	63186
39	71399
40	66389
41	60610
42	36650
43	27401
44	30700
47	26789
48	91429

Table S1

長分布を調べる

Length	Total
18nt	55702
19nt	106074
20nt	521188
21nt	267851
22nt	300121
23nt	167156
24nt	138656
25nt	174267
26nt	197271
27nt	339876
28nt	337492
29nt	132016
30nt	58901
31nt	69938
32nt	1551638
33nt	6529060
34nt	141067
35nt	72969
36nt	73926
37nt	99344
38nt	63419
39nt	71469
40nt	66514
41nt	61526
42nt	37604
43nt	27361
44nt	29035

原著論文と似た結果になっていることがわかる

Nie *et al.*, *BMC Genomics*, **14**: 1, 2013

Step4：カイコゲノムへのマッピング

Step4. カイコゲノムへのマッピングおよびカウントデータ取得:

マップしたい前処理後のFASTQファイル名("SRR609266_p.fastq.gz")およびその任意のサンプル名を記述した srp016842_samplename.txtを作業ディレクトリに保存したうえで、下記を実行します。

Step2でダウンロードしたintegretedseq.faへマッピングしています。

basic alignerの一つであるBowtieを内部的に用いており、ここではマッピング時のオプションを"-m 1 -v 0"とし、「ミスマッチを許容せず1ヶ所にのみマップされるもの」をレポートしています。ミスマッチを許容していないため、--best --strataというオプションは事実上意味をなさないためにつけていません。QuasRのマニュアル中のようにalignmentParameterオプションは特に指定せず(デフォルト)、50ヶ所にマップされるリードまでをレポートする"maxHits=50"オプションをつけるという思想もあります。

マシンパワーにもよりますが、20分程度で終わると思います。

マップ後 | カウント情報取得 | ゲノム | QuasR(Lerch_XXX)の記述内容と基本的に同じです。

```
in_f1 <- "srp016842_samplename.txt"     #入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "integretedseq.fa"             #入力ファイル名を指定してin_f2に格納(リファレンス配列)
out_f1 <- "srp016842_QC.pdf"            #出力ファイル名を指定して
out_f2 <- "srp016842_other_info1.txt"   #出力ファイル名を指定して
param_mapping <- "-m 1 -v 0"            #マッピング時のオプション

#必要なパッケージをロード
library(QuasR)                          #パッケージの読み込み
library(GenomicRanges)                  #パッケージの読み込み

#マッピングおよびQCレポート用ファイル作成
time_s <- proc.time()                   #計算時間を計測するため
out <- qAlign(in_f1, in_f2,             #マッピングを行うqAlign関数を実行した結果をoutに格納
      alignmentParameter=param_mapping)#マッピングを行うqAlign関数を実行した結果をoutに格納
time_e <- proc.time()                   #計算時間を計測するため
qQCReport(out, pdfFilename=out_f1)      #QCレポート結果をファイルに保存
```

srp016842_samplename.txt

FileName	SampleName
SRR609266_p.fastq.gz	sRNA

圧縮ファイルの状態でも指定可能
20分程度で終わります

Step4：カイコゲノムへのマッピング

Step4. カイコゲノムへのマッピングおよびカウントデータ取得:

マップしたい前処理後のFASTQファイル名("SRR609266_p.fastq.gz")およびその任意のサンプル名を記述した srp016842_samplename.txtを作業ディレクトリに保存したうえで、下記を実行します。
Step2でダウンロードしたintegretedseq.faへマッピングしています。
basic alignerの一つであるBowtieを内部的に用いており、ここではマッピング時のオプションを"-m 1 -v 0"とし、「ミスマッチを許容せず1ヶ所にのみマップされるもの」をレポートしています。ミスマッチを許容していないため、--best --strataというオプションは事実上意味をなさないためにつけていません。QuasRのマニュアル中のようにalignmentParameterオプションは特に指定せず(デフォルト)、50ヶ所にマップされるリードまでをレポートする "maxHits=50"オプションをつけるという思想もあります。
マシンパワーにもよりますが、20分程度で終わると思います。
マップ後 | カウント情報取得 | ゲノム | QuasR(Lerch_XXX)の記述

```
in_f1 <- "srp016842_samplename.txt"      #入力ファイ
in_f2 <- "int                            ァイ
out_f1 <- "sr                            ァイ
out_f2 <- "sr                            ァイ
param_mapping                            ング

#必要なパッケー
library(QuasR                            ージ
library(Genom                            ージ

#マッピングおよ
time_s <- pro                            間を
out <- qAlign                            ング
      alignme                            ング
time_e <- pro                            間を
qQCReport(out                            ート

#後処理(マップ                           域に
tmpfname <- o                            ル名
tmpsname <- out@alignments[,2]           #サンプル名
for(i in 1:length(tmpfname)){            #サンプル数
  k <- readGAlignments(tmpfname[i])      #BAM形式フ
```

切り取り(T)
コピー(C)
貼り付け
すべて選択(A)
印刷(I)...
印刷プレビュー(N)...
Bing でマップ
Bing で翻訳
Google で検索
電子メール (Windows Live Hotmail)
すべてのアクセラレータ
Send to OneNote

コピペ

R Console

```
    xtabs

以下のオブジェクトはマスクされています (from 'package:base') :

    anyDuplicated, append, as.data.frame, as.vector, cbind,
    colnames, duplicated, eval, evalq, Filter, Find, get, intersect,
    is.unsorted, lapply, Map, mapply, match, mget, order, paste,
    pmax, pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce,
    rep.int, rownames, sapply, setdiff, sort, table, tapply, union,
    unique, unlist

要求されたパッケージ IRanges をロード中です
要求されたパッケージ XVector をロード中です
要求されたパッケージ Rbowtie をロード中です
> library(GenomicRanges)                    #パッケージの読み込み
>
> #マッピングおよびQCレポート用ファイル作成
> time_s <- proc.time()                     #計算時間を計測するため
> out <- qAlign(in_f1, in_f2,               #マッピングを行うqAlign関数$
+       alignmentParameter=param_mapping)#マッピングを行うqAlign関数$
Creating .fai file for: C:/Users/kadota/Desktop/SRP16842/integretedse$
```

無事マッピングが終了すると、指定した2つのファイルが生成されて

Step4：カイコゲノムへのマッピング

無事マッピングが終了すると、指定した2つのファイルが生成されているはずです。

1. QCレポートファイル(srp016842_QC.pdf): Quality Controlレポートです。よく利用されるFastQCのようなものです。
2. その他の各種情報ファイル(srp016842_other_info1.txt): 論文作成時に必要な、マッピング時に用いたオプション情報、マップされたリード数、Rおよび用いたパッケージのバージョン情報などを含みます。

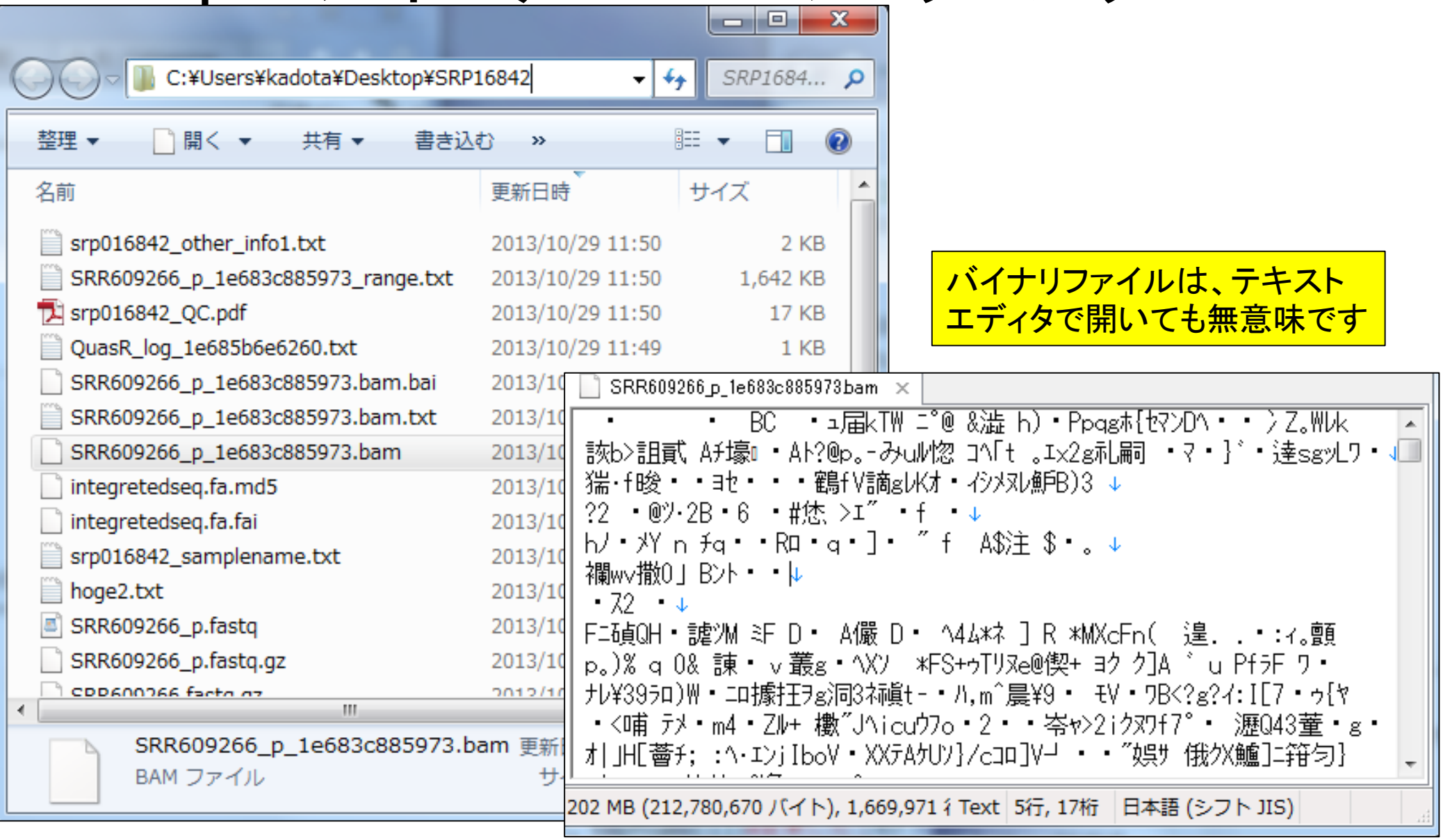
この他にも様々なファイルが生成されます。例えば、マッピング後に得られるBAM形式ファイルは、"SRR609266_p_XXXXXX.bam"というファイル名で作業ディレクトリ上に自動で生成されます (例: SRR609266_p_fa03ced5b37.bam; 約200Mb)。ここで、XXXXXXはランダムな文字列からなります。理由は、同じ入力ファイルを異なるパラメータやリファレンス配列にマッピングしたときに、上書きしてしまう恐れがあるためです。また、"SRR609266_p_XXXXXX_range.txt"というファイルも生成されます(例: SRR609266_p_fa03ced5b37_range.txt)。これは、マップされたリードの和集合領域(オーバーラップ領域をまとめたもの)をリストアップし、その領域に基づいてカウント情報を取得したものです。BEDTools(Quinlan and Hall, Bioinformatics, 2010)のmergeBEDというプログラムと同じような結果を得ているという理解でかまいません。

- SRP016842: Nie et al., BMC Genomics, 2013
- SRAdb: Zhu et al., BMC Bioinformatics, 2013
- QuasR: 原著論文はまだみたいです
- Bowtie: Langmead et al., Genome Biol., 2009
- GenomicFeatures: Lawrence et al., PLoS Comput. Biol., 2013

基本的なマッピング結果はバイナリのBAM形式ファイルに保存されます。

```
> cat("1. Computation time for mapping (in second).\n")#計算時間を表$
> time_e - time_s                          #計算時間を表示(一番右側の数$
> cat("\n\n2. Options used for mapping.\n")#マッピングに用いたオプシ$
> out@alignmentParameter                   #マッピングに用いたオプショ$
> cat("\n\n3. Alignment statistics.\n")     #マッピング結果(alignment st$
> alignmentStats(out)                      #マッピング結果(alignment st$
> cat("\n\n4. Session info.\n")             #解析に用いたRや各種パッケー$
> sessionInfo()                            #解析に用いたRや各種パッケー$
> sink()                                   #書き込み終了の指令
>
```

Step4：カイコゲノムへのマッピング



バイナリファイルは、テキスト
エディタで開いても無意味です

Step4：カイコゲノムへのマッピング

・RでBAM形式ファイルを読み込めます
・BAM → BED形式への変換もできます

R Console

```
> hoge <- readGAlignments("SRR609266_p_1e683c885973.bam")
> hoge
GAlignments with 947184 alignments and 0 metadata columns:
           seqnames strand       cigar    qwidth     start       end     width      ngap
              <Rle>  <Rle> <character> <integer> <integer> <integer> <integer> <integer>
       [1]     chr1      +         26M        26      3609      3634        26         0
       [2]     chr1      +         26M        26      3609      3634        26         0
       [3]     chr1      -         28M        28      5423      5450        28         0
       [4]     chr1      -         28M        28      5423      5450        28         0
       [5]     chr1      -         26M        26     53389     53414        26         0
       ...      ...    ...         ...       ...       ...       ...       ...       ...
  [947180]    chr28      +         20M        20  12349897  12349916        20         0
  [947181]    chr28      +         20M        20  12349897  12349916        20         0
  [947182]    chr28      +         20M        20  12349897  12349916        20         0
  [947183]    chr28      +         20M        20  12349904  12349923        20         0
  [947184]    chr28      +         21M        21  12349908  12349928        21         0
  ---
  seqlengths:
       chr1     chr2     chr3     chr4     chr5 ...    chr24    chr25    chr26    chr27    chr28
   22395194 10467956 18213042 20915959 20918634 ... 18494581 16676770 12282741 14467522 12350153
> 
```

- イントロ｜ファイル形式の変換｜について(last modified 2013/09/30) NEW
- イントロ｜ファイル形式の変換｜BAM --> BED(last modified 2013/10/25) NEW
- イントロ｜ファイル形式の変換｜FASTQ --> FASTA(last modified 2013/06/17)
- イントロ｜ファイル形式の変換｜qseq --> FASTA(last modified 2013/06/17)
- イントロ｜ファイル形式の変換｜qseq --> Illumina FASTQ(last modified 2013/06/17)
- イントロ｜ファイル形式の変換｜qseq --> Sanger FASTQ(last modified 2013/08/19)

Step4：カイコゲノムへのマッピング

R Console

```
> hoge <- readGAlignments("SRR609266_p_1e683c885973.bam")
> hoge
GAlignments with 947184 alignments and 0 metadata columns:
            seqnames strand       cigar    qwidth     start       end     width      ngap
               <Rle>  <Rle> <character> <integer> <integer> <integer> <integer> <integer>
       [1]     chr1      +         26M        26      3609      3634        26         0
       [2]     chr1      +         26M        26      3609      3634        26         0
       [3]     chr1      -         28M        28      5423      5450        28         0
       [4]     chr1      -         28M        28      5423      5450        28         0
       [5]     chr1      -         26M        26     53389
       ...      ...    ...         ...       ...       ...
  [947180]     chr28     +         20M        20  12349897  12
  [947181]     chr28     +         20M        20  12349897  12349916        20         0
  [947182]     chr28     +         20M        20  12349897  12349916        20         0
  [947183]     chr28     +         20M        20  12349904  12349923        20         0
  [947184]     chr28     +         21M        21  12349908  12349928        21         0
  ---
  seqlengths:
       chr1      chr2                                                    27     chr28
   22395194  10467956                                                    22  12350153
> |
```

マップされたリードの和集合領域同定
およびリード数のカウントもできます

SRR609266_p_1e683c885973_range.txt

seqnames	start	end	width	strand	sRNA
chr1	3609	3634	26	+	2
chr1	78269	78302	34	+	1
chr1	79373	79398	26	+	1
...					

Step4：カイコゲノムへのマッピング

R Console

```
> hoge <- readGAlignments("SRR6092
> hoge
GAlignments with 947184 alignments
            seqnames strand       c
               <Rle>  <Rle> <charac
        [1]      chr1      +
        [2]      chr1      +
        [3]      chr1      -
        [4]      chr1      -          28M        28     5423     5450       28         0
        [5]      chr1      -          26M        26    53389    53414       26         0
        ...       ...    ...          ...       ...      ...      ...      ...       ...
   [947180]     chr28      +          20M        20 12349897 12349916       20         0
   [947181]     chr28      +          20M        20 12349897 12349916       20         0
   [947182]     chr28      +          20M        20 12349897 12349916       20         0
   [947183]     chr28      +          20M        20 12349904 12349923       20         0
   [947184]     chr28      +          21M        21 12349908 12349928       21         0
   ---
   seqlengths:
        chr1      chr2      chr3      chr4      chr5 ...     chr24     chr25     chr26     chr27     chr28
    22395194  10467956  18213042  20915959  20918634 ...  18494581  16676770  12282741  14467522  12350153
> |
```

SRR609266_p_1e683c885973_range.txt

chr28	12349103	12349125	23	+	2
chr28	12349893	12349928	36	+	7
chr28	213283	213303	21	−	1
chr28	274128	274149	22	−	1

マップされたリードの和集合領域同定
およびリード数のカウントもできます

まとめ

- *SRAdb* (Zhu et al., *BMC Bioinformatics*, **14**: 19, 2013)
  - 公共DBからのRNA-seqデータ（FASTQファイル）取得
- *QuasR* (Lerchら, unpublished)
  - リファレンス配列（ゲノム or トランスクリプトーム）へのマッピング
    - Bowtie (Langmeadら, 2009) or SpliceMap (Auら, 2010)を選択可能
    - 出力はBAM形式ファイル、QCレポートも
  - 遺伝子アノテーション情報をもとにカウントデータ取得
    - *GenomicFeatures* (Lawrenceら, 2013)から得られるTranscriptDbオブジェクトを利用
    - UCSC known genesやEnsembl genesのカウントデータなど
- *TCC* (Sun et al., *BMC Bioinformatics*, **14**: 219, 2013)
  - 内部的に*edgeR* (Robinsonら, 2010)や*DESeq* (Anders, 2010)などを用いて頑健な発現変動解析を実行

(アセンブル以外の)一通りのRNA-seq解析は、Linuxコマンド抜きで可能です

謝辞

共同研究者

清水 謙多郎　先生（東京大学・大学院農学生命科学研究科）

西山 智明　先生（金沢大学・学際科学実験センター）

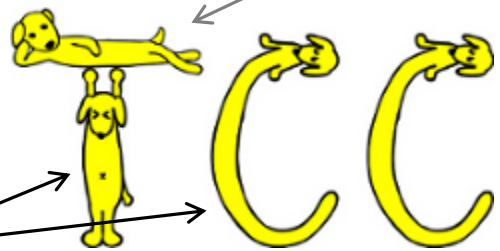
孫 建強　氏（東京大学・大学院農学生命科学研究科・大学院生）

グラント

- 基盤研究(C)（H24-26年度）:「シークエンスに基づく比較トランスクリプトーム解析のためのガイドライン構築」(代表)
- 新学術領域研究(研究領域提案型)(H22年度-):「非モデル生物におけるゲノム解析法の確立」(分担；研究代表者：西山智明)

挿絵やTCCのロゴなど

（妻の）門田 雅世さま作



（有能な秘書の）三浦 文さま作