RNA-Seq演習

千葉大学真菌医学研究センター
高橋弘喜
hiroki.takahashi@chiba-u.jp

演習の内容

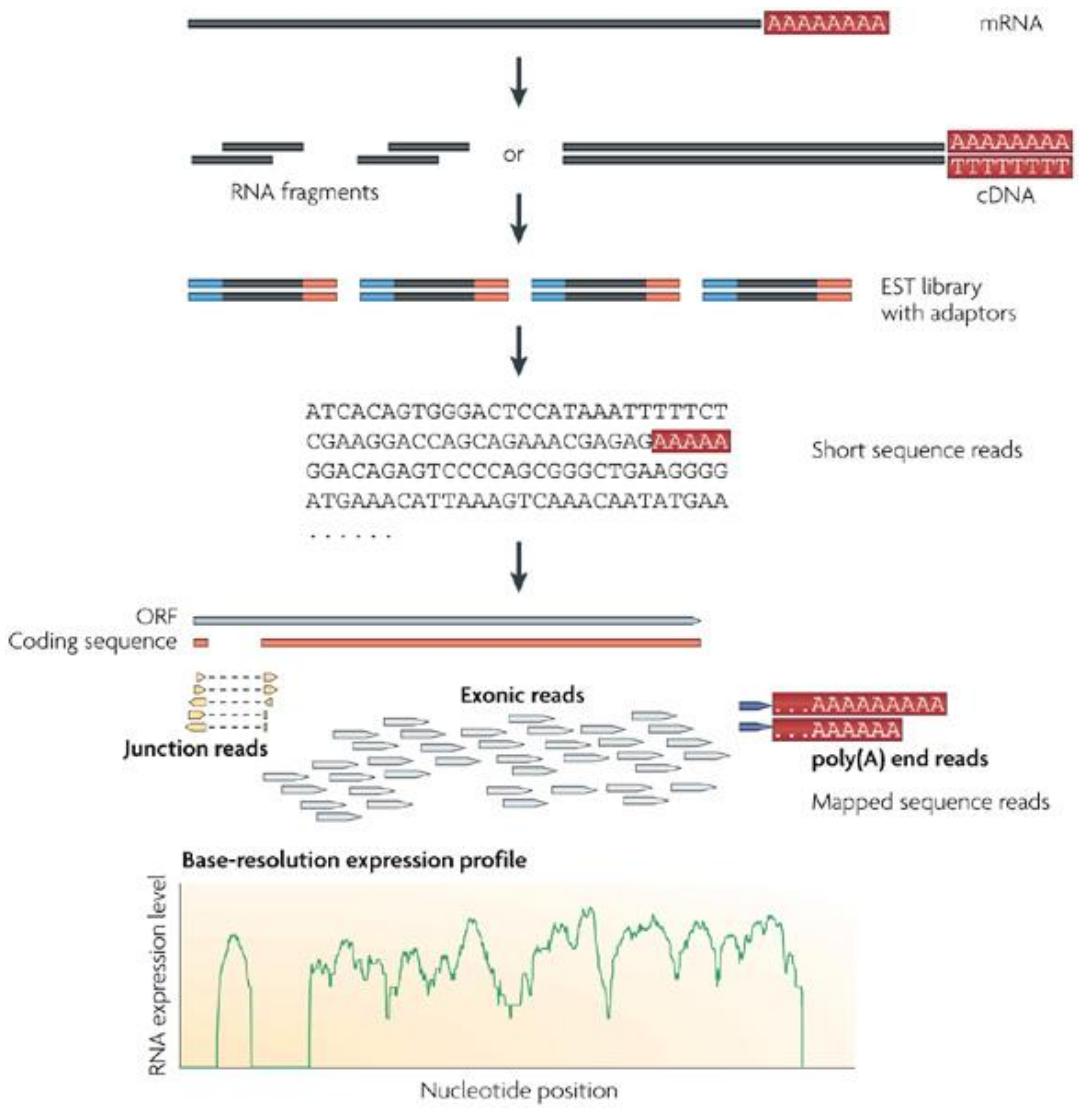
- テストデータを用いて、リードのマッピングから発現量の算出、発現変化を示す遺伝子のリストアップまでの一連の解析を実施する。

- テストデータ（再度ダウンロード願います）

http://bioinfo.pf.chiba-u.jp/enshu_20151120/enshu_data.zip

名前	種類	圧縮サイズ	パスワー...	サイズ
annot	ファイル フォルダー			
fastq	ファイル フォルダー			
results	ファイル フォルダー			
cuffdiff.sh	SH ファイル	1 KB	無	
enshu.R	R ファイル	1 KB	無	
memo	ファイル	1 KB	無	
tophat.sh	SH ファイル	1 KB	無	

RNA-Seq



RNA抽出

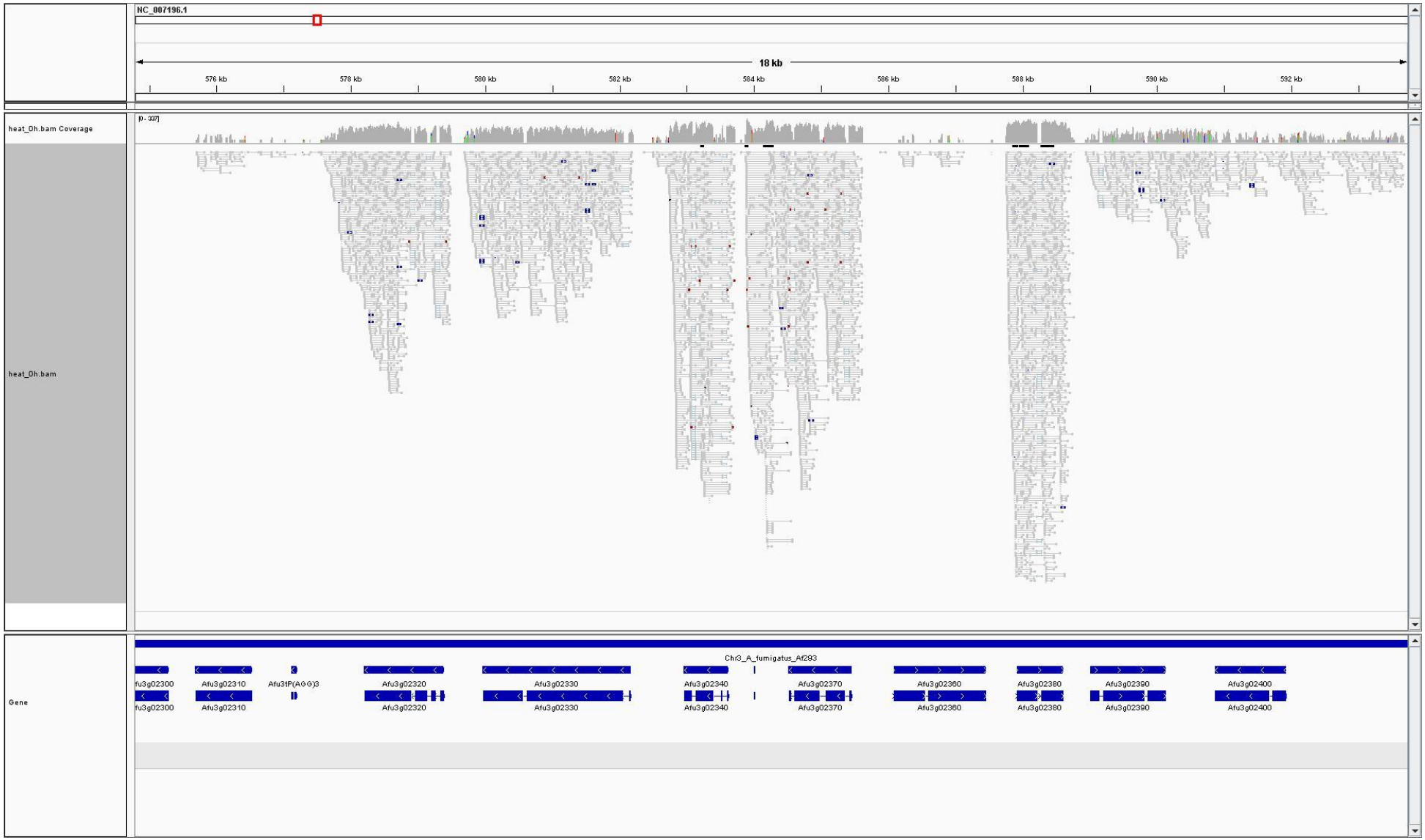
ライブラリー作製

シーケンス

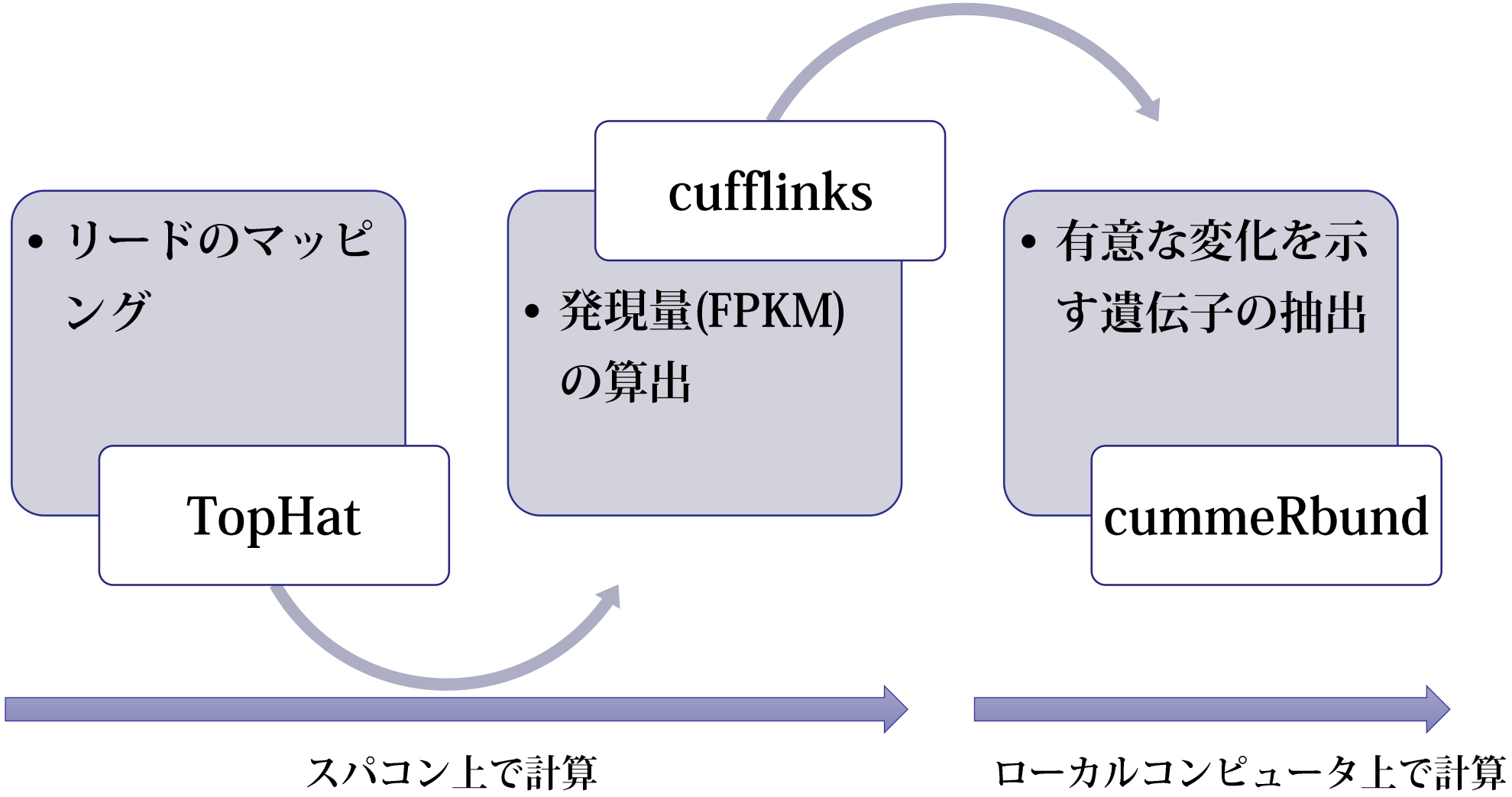
データ解析

Wang et al. Nature Reviews Genetics 10, 57-63 (January 2009)

結果例


NC_007196.1
18 kb
576 kb
578 kb
580 kb
582 kb
584 kb
586 kb
588 kb
590 kb
592 kb
heat_Oh.bam Coverage
heat_Oh.bam
Gene
Chr3_A_fumigatus_Af293
Afu3g02300
Afu3g02310
Afu3tP(AGG)3
Afu3g02320
Afu3g02330
Afu3g02340
Afu3g02370
Afu3g02360
Afu3g02380
Afu3g02390
Afu3g02400

本日の内容


• リードのマッピング
TopHat
cufflinks
• 発現量(FPKM)の算出
• 有意な変化を示す遺伝子の抽出
cummeRbund
スパコン上で計算
ローカルコンピュータ上で計算

スパコン使用方法(イメージ)

①ゲートウェイノード(gw.ddbj.nig.ac.jp)にログインする
②qloginを実行しインタラクティブノードにログインする
③qloginしたホストからジョブをUGEに投入する
④UGEは負荷の低いノードでジョブを実行する
⑤ジョブ実行結果をlustreのホームディレクトリに出力する
⑥ジョブ実行結果を確認する

fastq

```
1 @SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
2 GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
3 +SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
4 IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

- 1行目:'@'に続き配列ID。
- 2行目:配列の文字列
- 3行目:‘+’に続いて、配列IDか、改行
- 4行目:asciiコードで表現したquality value

fastq

33:! 34:" 35:# 36:$ 37:% 38:& 39:' 40:(41:)
42:* 43:+ 44:, 45:- 46:. 47:/ 48:0 49:1 50:2
51:3 52:4 53:5 54:6 55:7 56:8 **57:9** 58:: 59:;
60:< 61:= 62:> 63:? 64:@ 65:A 66:B **67:C** 68:D
69:E 70:F **71:G** 72:H **73:I** 74:J 75:K 76:L 77:M
78:N 79:O 80:P 81:Q 82:R 83:S 84:T 85:U 86:V
87:W 88:X 89:Y 90:Z 91:[92:¥ 93:] 94:^ 95:_
96:` 97:a 98:b 99:c 100:d 101:e 102:f 103:g
104:h 105:i 106:j 107:k 108:l 109:m 110:n 111:o
112:p 113:q 114:r 115:s 116:t 117:u 118:v 119:w
120:x 121:y 122:z 123:{ 124:| 125:} 126:~

fastq

```
2 GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
4 IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

S　Q　P（塩基の信頼性）

I:73→40→1.0×10^{-4}

9:57→24→3.9×10^{-3}

G:71→38→1.6×10^{-4}

C:67→34→4.0×10^{-4}

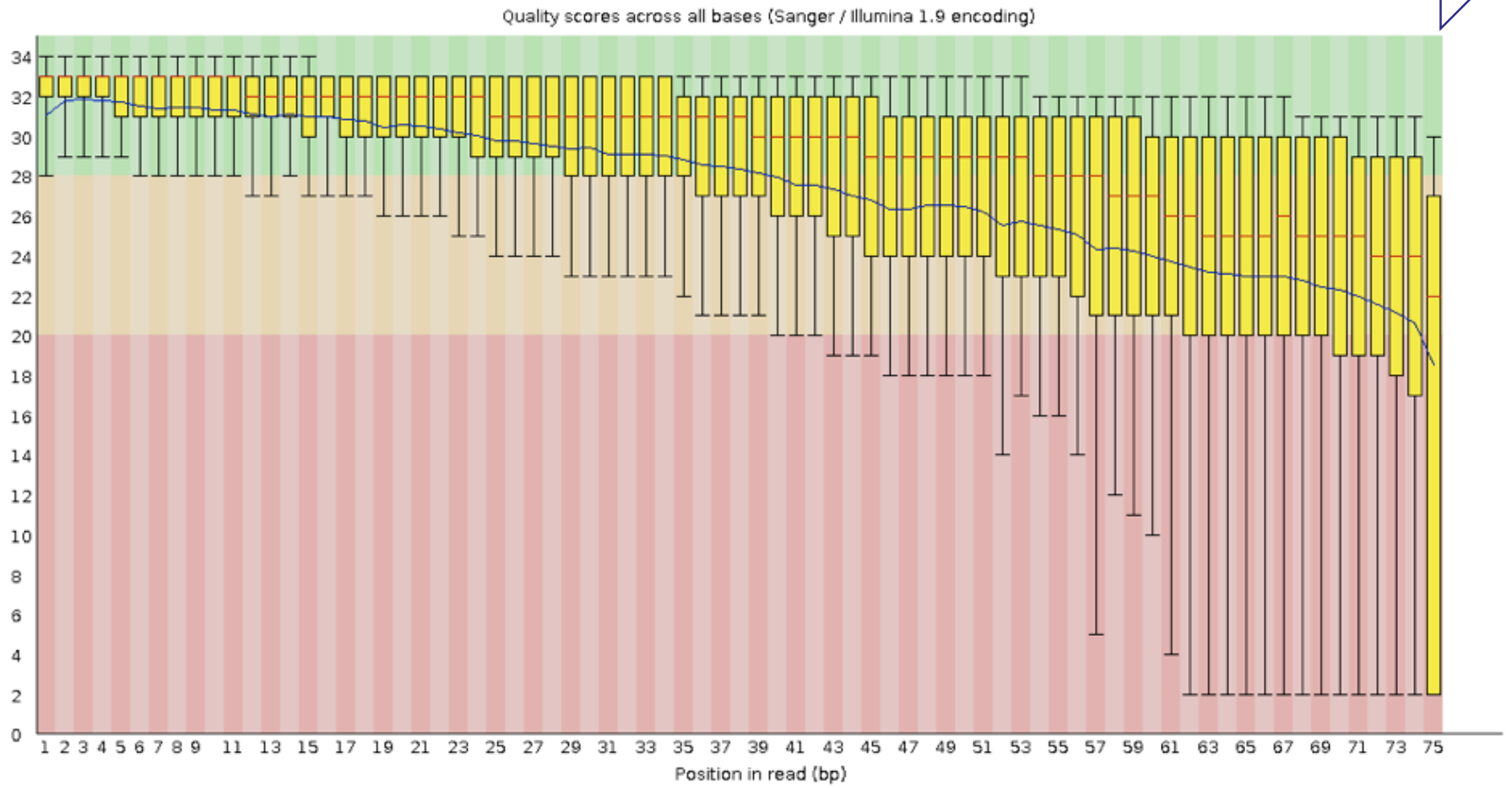
$$Q = -10\log_{10}p$$

$$40 = -10\log_{10} P$$

$$P = 10^{-4}$$

データクオリティー

シーケンスが進むにつれて信頼性が落ちていく



FastQC http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

テストデータ

DRASearch　Send Feedback　Search Home　DRA Home

SRA051410 FTP

Submission Detail	
Alias	S. cerevisiae CENPK RNA-seq
Submission ID	
Submission Date	2012-04-04
Center Name	Chalmers University of Technology
Lab Name	

Navigation			
Study	SRP012047		
Experiment	SRX135198	FASTQ	SRA
	SRX135710	FASTQ	SRA
	SRX135711	FASTQ	SRA
	SRX135712	FASTQ	SRA
	SRX135713	FASTQ	SRA
Sample	SRS307298		
	SRS308058		
Run	SRR453566	FASTQ	SRA
	SRR453567	FASTQ	SRA
	SRR453568	FASTQ	SRA
	SRR453569	FASTQ	SRA
	SRR453570	FASTQ	SRA
	SRR453571	FASTQ	SRA
	SRR453572	FASTQ	SRA
	SRR453573	FASTQ	SRA
	SRR453574	FASTQ	SRA
	SRR453575	FASTQ	SRA
	SRR453576	FASTQ	SRA
	SRR453577	FASTQ	SRA
	SRR453578	FASTQ	SRA

http://trace.ddbj.nig.ac.jp/DRASearch/

Ogasawara et al. *Nucleic Acids Res.* 2014, 42: D44-D49.

テストデータ

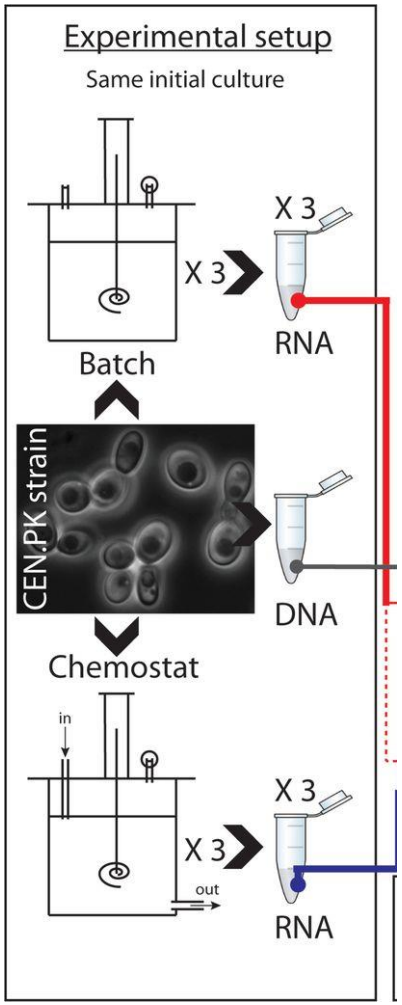
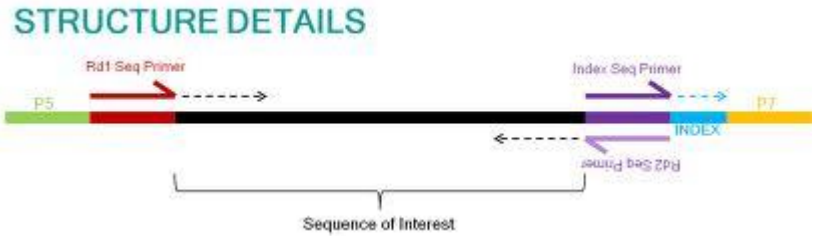


Figure 1

Illumina HiSeq2000
100bp PE

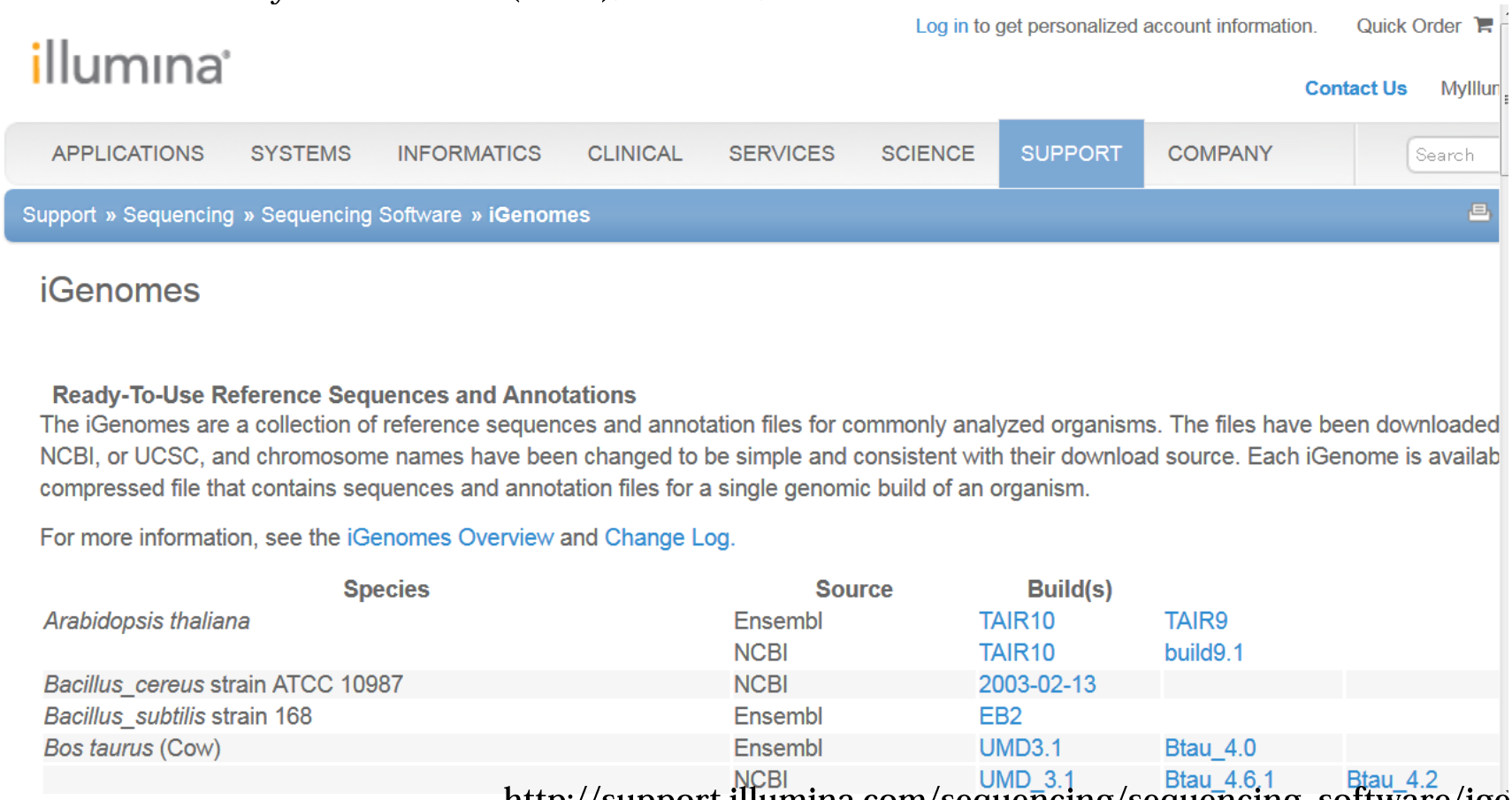


サンプル	Run	リード数	テストデータ
batch1	SRR453566	5,725,730	10,000
batch2	SRR453567	7,615,732	
batch3	SRR453568	5,565,734	
chemo1	SRR453569	4,032,514	
chemo2	SRR453570	6,745,975	
chemo3	SRR453571	6,163,396	

Nookaew et al. *Nucl. Acids Res.* (2012) 40 (20): 10084-10097.

アノテーションファイル

- iGenomesから取得
  - *Saccharomyces cerevisiae* (Yeast), Ensembl, R64-1-1



illumina

Log in to get personalized account information. Quick Order

Contact Us MyIllur

APPLICATIONS SYSTEMS INFORMATICS CLINICAL SERVICES SCIENCE SUPPORT COMPANY Search

Support » Sequencing » Sequencing Software » iGenomes

iGenomes

Ready-To-Use Reference Sequences and Annotations

The iGenomes are a collection of reference sequences and annotation files for commonly analyzed organisms. The files have been downloaded NCBI, or UCSC, and chromosome names have been changed to be simple and consistent with their download source. Each iGenome is availab compressed file that contains sequences and annotation files for a single genomic build of an organism.

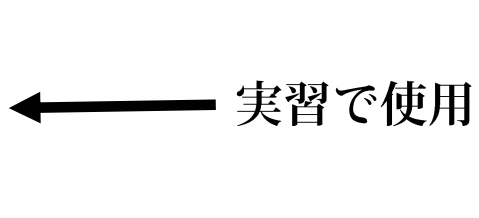
For more information, see the iGenomes Overview and Change Log.

Species	Source	Build(s)		
Arabidopsis thaliana	Ensembl	TAIR10	TAIR9	
	NCBI	TAIR10	build9.1	
Bacillus_cereus strain ATCC 10987	NCBI	2003-02-13		
Bacillus_subtilis strain 168	Ensembl	EB2		
Bos taurus (Cow)	Ensembl	UMD3.1	Btau_4.0	
	NCBI	UMD_3.1	Btau_4.6.1	Btau_4.2

http://support.illumina.com/sequencing/sequencing_software/igenome.html

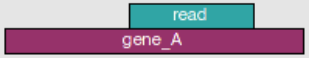
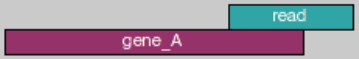
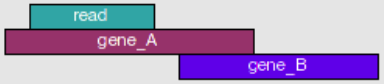
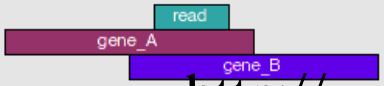
アノテーションファイル

- iGenomeでは、マッピングに必要なインデックスファイル、アノテーションファイルが用意されている。
  - BWAIndex
  - Bowtie2Index
  - BowtieIndex
  - genes.gtf



その他

- どのリードを転写物由来とするか

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

http://www-huber.embl.de/users/anders/HTSeq/doc/count.html

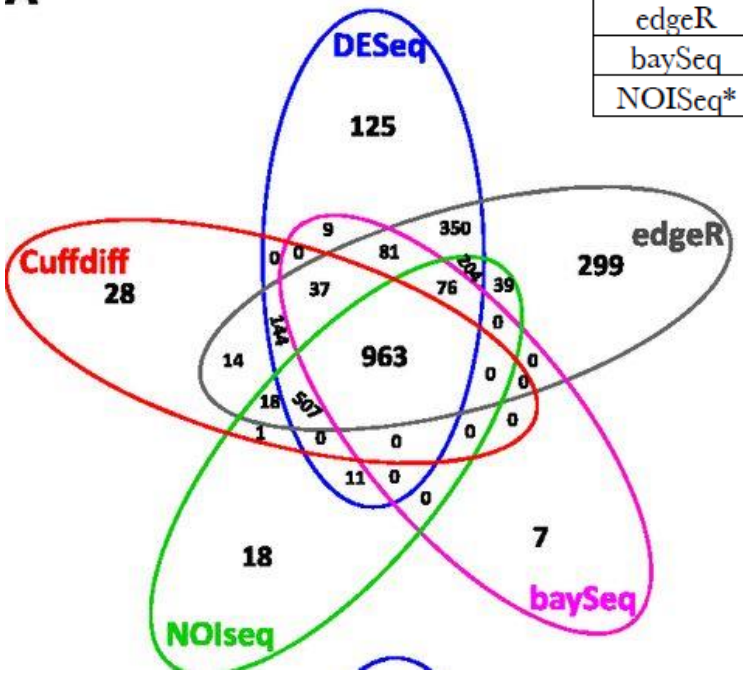
その他

- 解析手法（マッピング含む）による違い

Table S4 Number of DGE (Q-values < 10e-5) from different methods (for microarray the number DGE is 1603)

Method	Gsnap	N.Gsnap	Stampy	N.Stampy	TopHat	N.TopHat	De novo
Cufdiff	2061	2172	1712	1741	1671	1726	1623
DESeq	2690	2731	2507	2503	2412	2432	2197
edgeR	3087	3161	2732	2742	2649	2673	2385
baySeq	1785	1807	1173	1198	1092	1133	1175
NOISeq*	2097	2070	1837	1784	1804	1754	1595

A



Nookaew et al. *Nucl. Acids Res.* (2012) 40 (20): 10084-10097.

その他

- 別の解析法
  - HTSeq (http://www-huber.embl.de/HTSeq/doc/overview.html)
  - DESeq
    (http://bioconductor.org/packages/release/bioc/html/DESeq.html)
  - edgeR
    (http://bioconductor.org/packages/release/bioc/html/edgeR.html)
- Viewer
  - IGV (http://www.broadinstitute.org/igv/)
- データQC
  - FastQC
    (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)