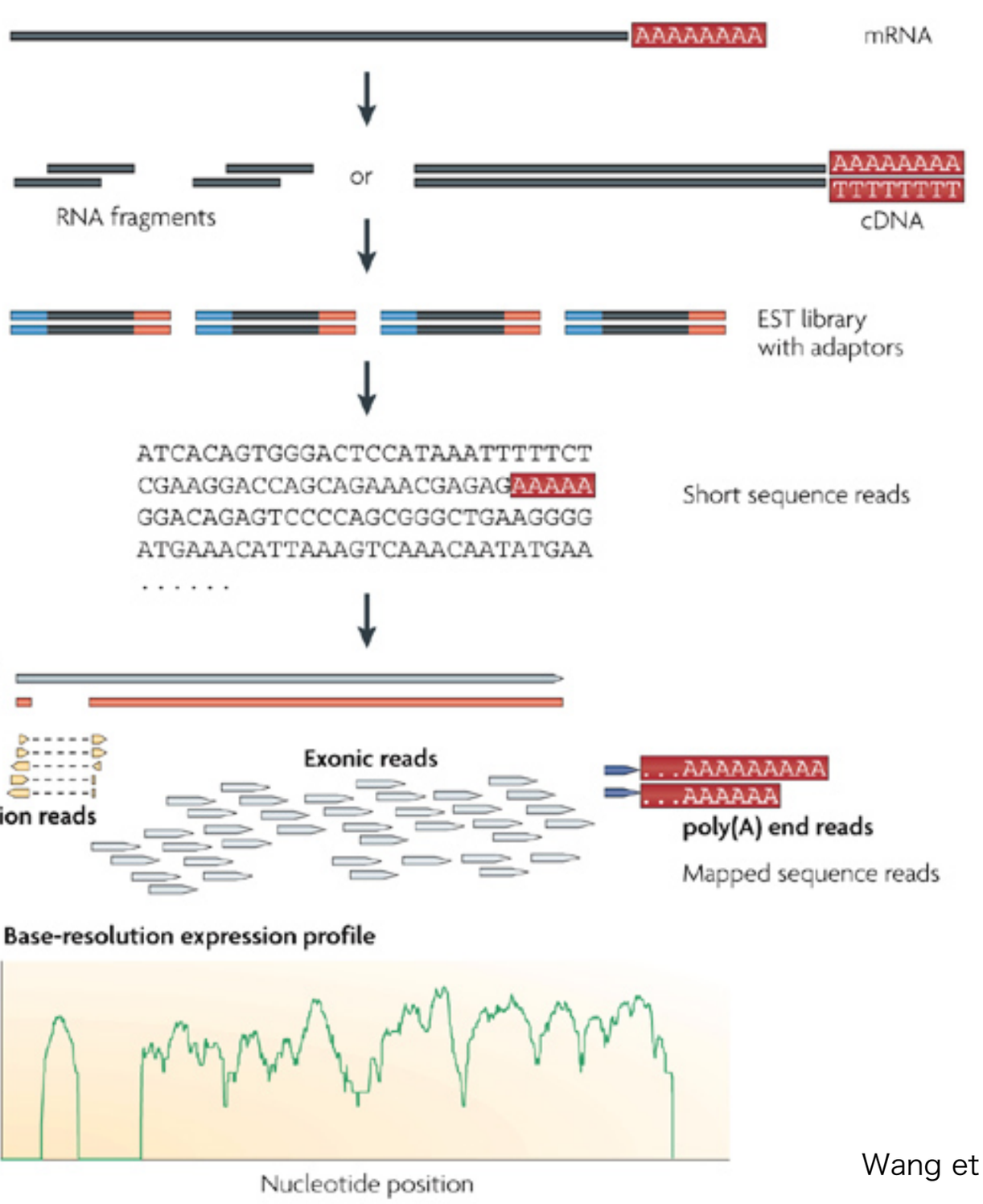
RNA-seq演習

千葉大学真菌医学研究センター
高橋　弘喜

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までの解析を遺伝研スパコン上でやってみる.

RNA-seq



RNA抽出

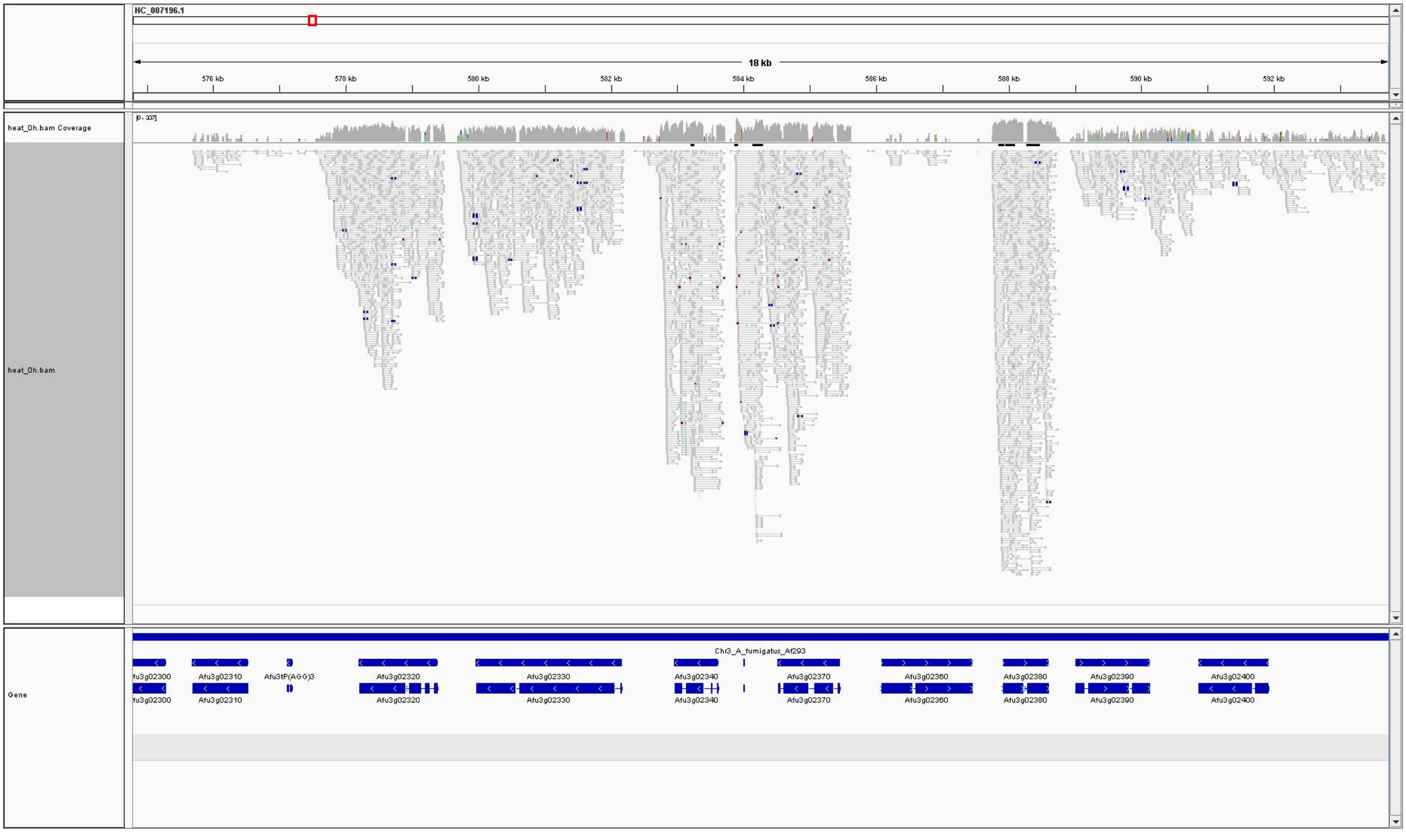
ライブラリー調整

シーケンス

データ解析

Wang et al., Nat Rev Genet 10(1):57-63, 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_0h.bam Coverage
[0 - 337]
heat_0h.bam
Gene
Chr3_A_fumigatus_Af293
fu3g02300
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
（TopHat2）

発現量の算出
（cufflinks）

有意な発現変化を示す遺伝子の探索
（cummeRbund）

スパコン上で計算

ローカルコンピューター上で計算

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

①

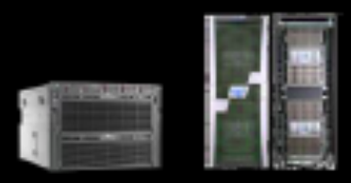
②

1台(active-stanby)
外部との通信専用

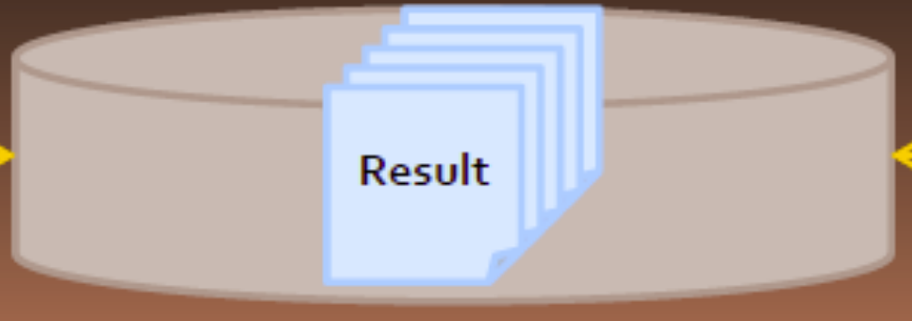
業務用8台、研究用8台
インタラクティブ専用

③

④



業務用81台、研究用208台
バッチ専用



⑥

⑤

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→1.0x10^{-4}
9:57→24→3.9x10^{-3}
G:71→38→1.6x10^{-4}
C:67→34→4.0x10^{-4}

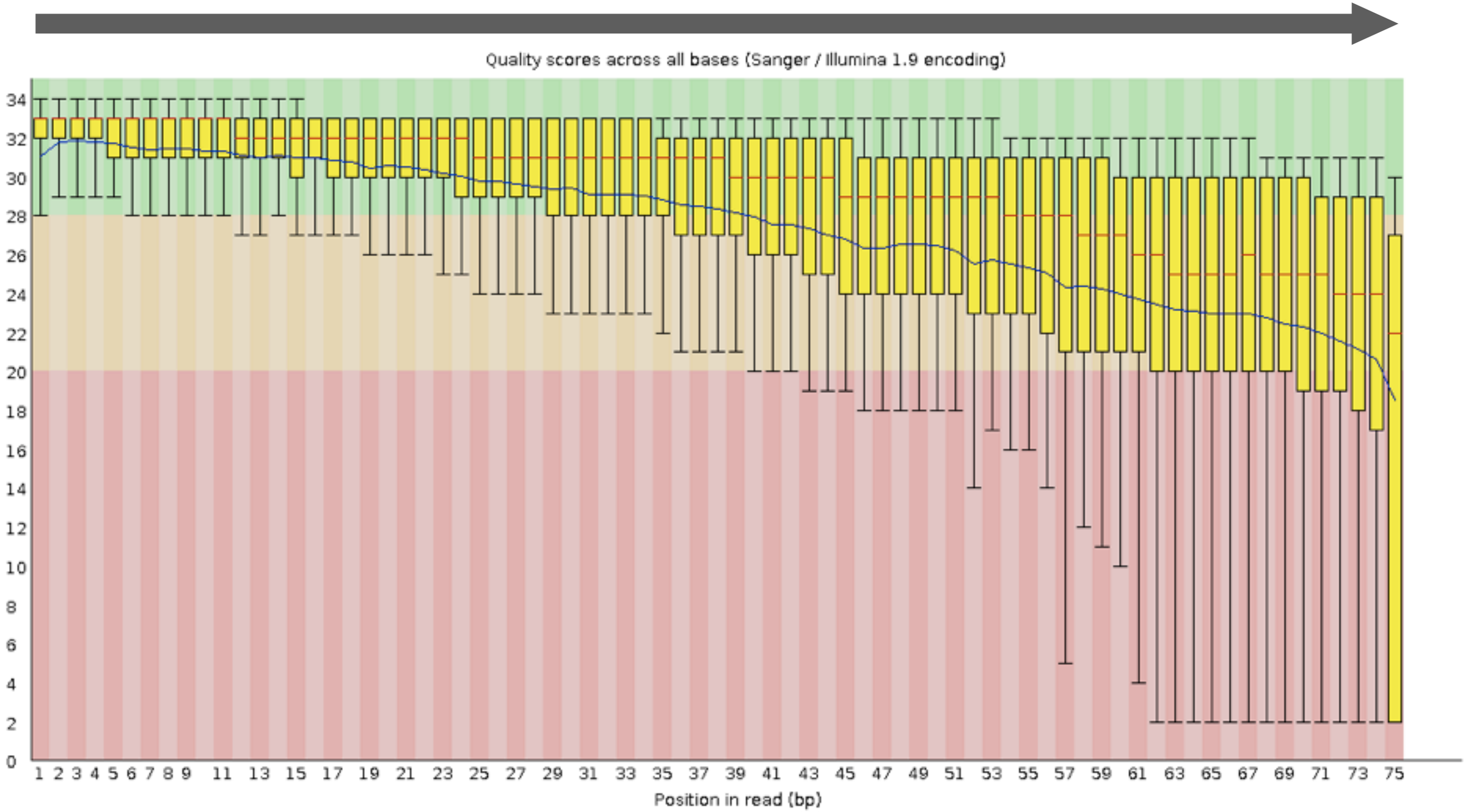
$$Q = -10\log_{10}p$$

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

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

データクオリティ

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/

sam/bam

- ＮＧＳデータをリファレンスへマッピングすると，sam形式（テキストファイル）で出力される.　（DNA-seq,RNA-seq, ChIP-seqなど）
- 11列のデータ（タブ区切り）で記載されている.

sam/bam

QNAME	Query template NAME
FLAG	bitwise FLAG
RNAME	References sequence NAME
POS	1- based leftmost mapping POSition
MAPQ	MAPping Quality
CIGAR	CIGAR String
RNEXT	Ref. name of the mate/next read
PNEXT	Position of the mate/next read
TLEN	observed Template LENgth
SEQ	segment SEQuence
QUAL	ASCII of Phred-scaled base QUALity+33

Sequence Alignment/Map Format Specification

例

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002        aaaAGATAA*GGATA
+r003      gcctaAGCTAA
+r004                  ATAGCT..............TCAGC
-r003                        ttagctTAGGC
-r001/2                                  CAGCGGCAT
```

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Sequence Alignment/Map Format Specification

sam⇔bam

- bamファイルは，samファイルをバイナリーに変換したもの．それにより，データサイズの圧縮，データへの高速なアクセスが可能になる.
- sam⇔bamは，samtoolsを使用することで相互に変換可能.

samtools

- sam→bam

  $ samtools view -bS test.sam > test.bam

- bam→sam

  $ samtools view -h test.bam > test.sam

- ソート（test.sort.bamが作成）

  $ samtools sort test.bam test.sort

- indexの作成（test.sort.bam.baiが作成）

  $ samtools index test.sort.bam

Li et al. Bioinformatics 25(16):2078-9, 2009

テストデータ

DRASearch

Send Feedback | Search Home | DRA Home

SRA051410 FTP

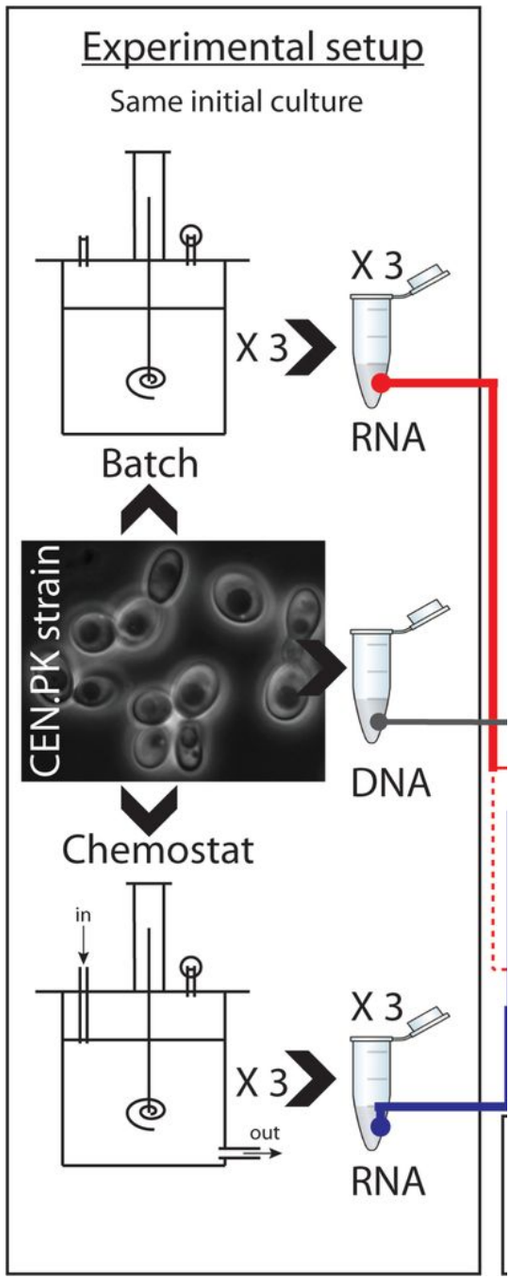
Submission Detail	
Alias	S. cerevisiae CENPK RNA-seq
Submission ID	
Submission Date	2012-04-04
Center Name	Chalmers University of Technology
Lab Name	

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

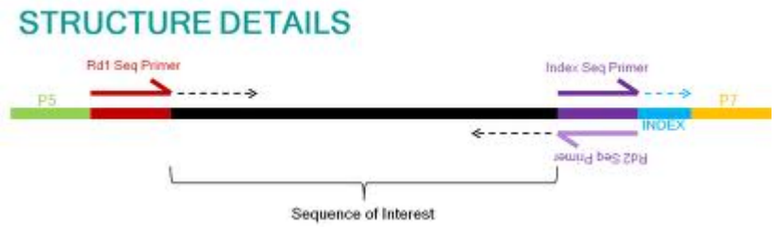
http://trace.ddbj.nig.ac.jp/DRASearch/

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

テストデータ



Illumina HiSeq2000
100bp PE

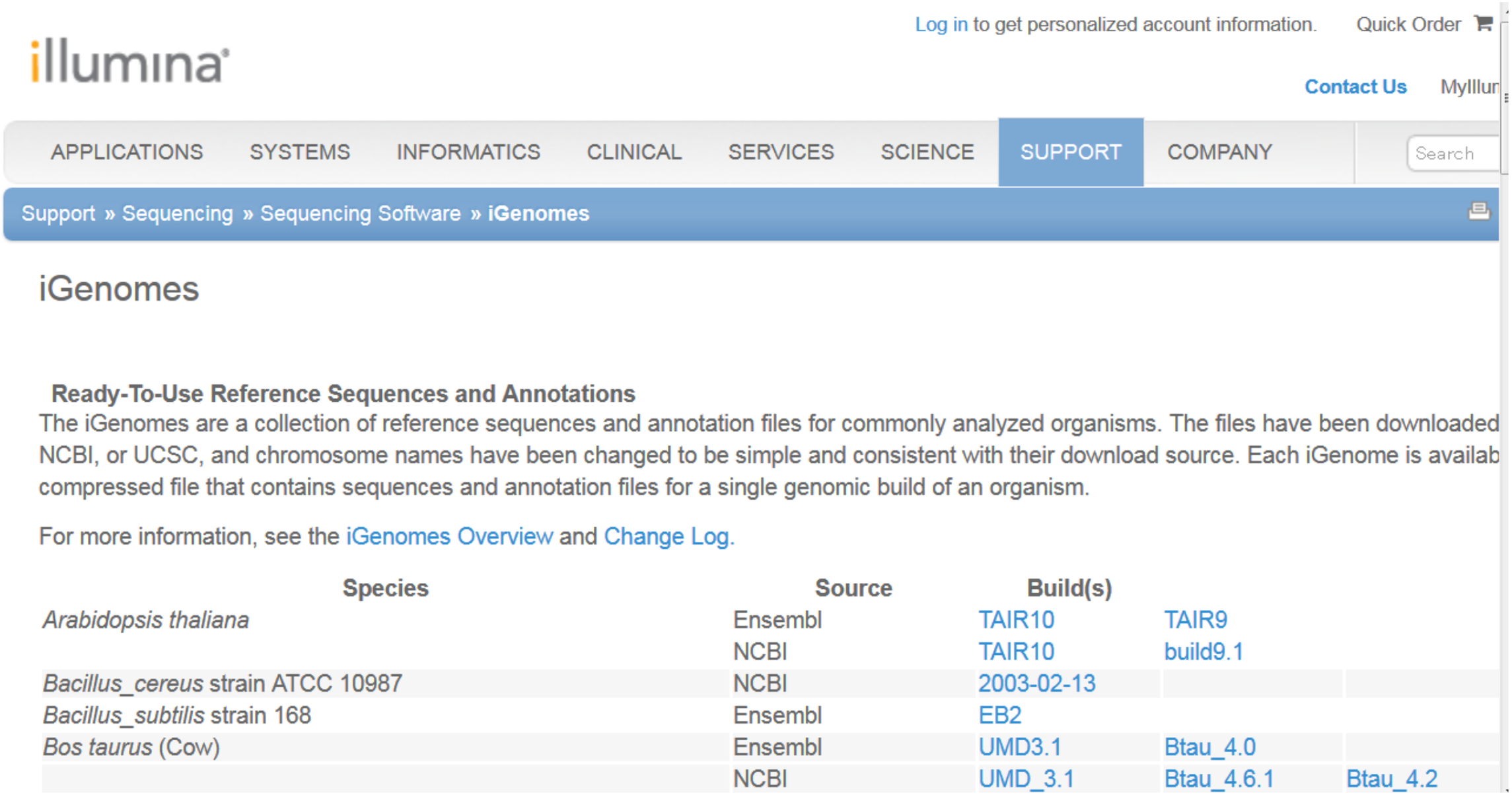


サンプル	Run	リード数
batch1	SRR453566	5,725,730
batch2	SRR453567	7,615,732
batch3	SRR453568	5,565,734
chemo1	SRR453569	4,032,514
chemo2	SRR453570	6,745,975
chemo3	SRR453571	6,163,396

Nookaew et al., Nucleic Acids Res 40(20):10084-97, 2012

アノテーションファイル

iGenomesから取得



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

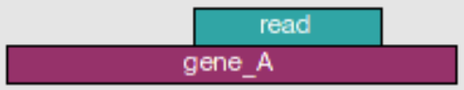
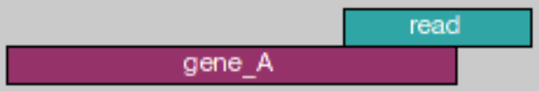
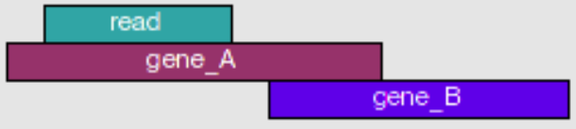
アノテーションファイル

- iGenomes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

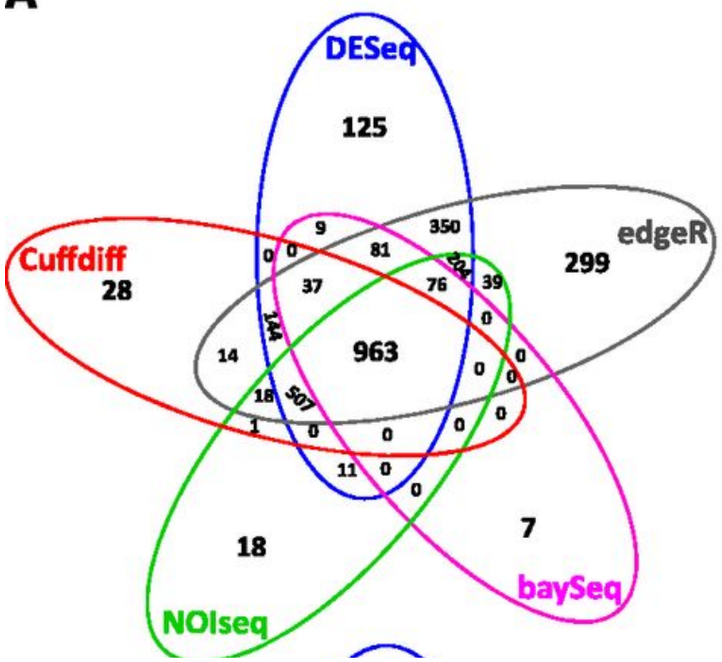
その他

- 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., Nucleic Acids Res 40(20):10084-97, 2012

その他

- 別の解析手法
  ーHTSeq
  ーDESeq2
  ーedgeR

- Viewer
  ーIGV

- データQC
  ーFastQC