

シーケンサー利用技術講習会 第10回 サンプルQC、RNAseqライブ ラリー作製/データ解析実習講習会

理化学研究所

ライフサイエンス技術基盤研究センター

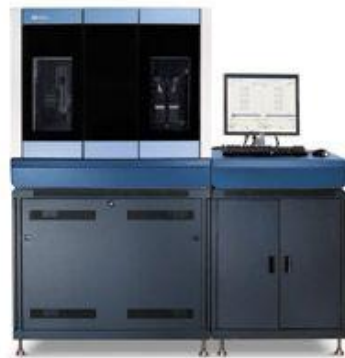
ゲノムネットワーク解析支援施設

田上 道平



次世代シーケンサー

Sequencer	File Format	Output(Max)	Read length
Illumina HiSeq2500	Fastq	600 Gb	100 bp
Life tech SOLiD	csfasta,qual	100 Gb	50 bp
Roche FLX	Sff	600 Mb	800 bp



Hiseq 2500/1500



Fastq data

```
@HWI-ST1394:58:H0B70ADXX:2:1101:4041:2089 1:N:0:  
TAAATGGTAGGGAAAGAGTGTAGGGAAAGAGTGAAGGAATAGCGTCGTGTTGGGTAAGAGTGAAGGGGTGTGGCTTTTAGTCATAGCTGTTTCCTGCTG  
+  
CCCFFFDHHHHBEIIBE3AAFHHDCEHGH??CGHGIGHIGFGIDGF7@FFHICCHHCE.=?E@CDFC99>@BBABCCCC@CDEECCCC+>>CCCCCC  
@HWI-ST1394:58:H0B70ADXX:2:1101:4204:2099 1:N:0:  
ATTTTTGTGGATGTATAGTTTATTTGTTGTTGGATTTGTTAGGATTTAAGTTTTTGGAGTATAATAGAGTTTAAAGATAAAAAGATTATTTTTGTA  
+  
CCCFFFFFFHHHJGHIIIIJGIJJJJJJHHHIJJGIIJJIJJGIIJJJJJHIJJJJGHHHHHHFFFFFFCDEEEEEEDDDDDDDDDDEEEDDD?4
```

@Header

TAAATGG.... (シーケンスで読まれた配列)

+

CCCFFFF... (クオリティースコア)

Quality Check for fastq data

- ソフトウェア

- FastQC

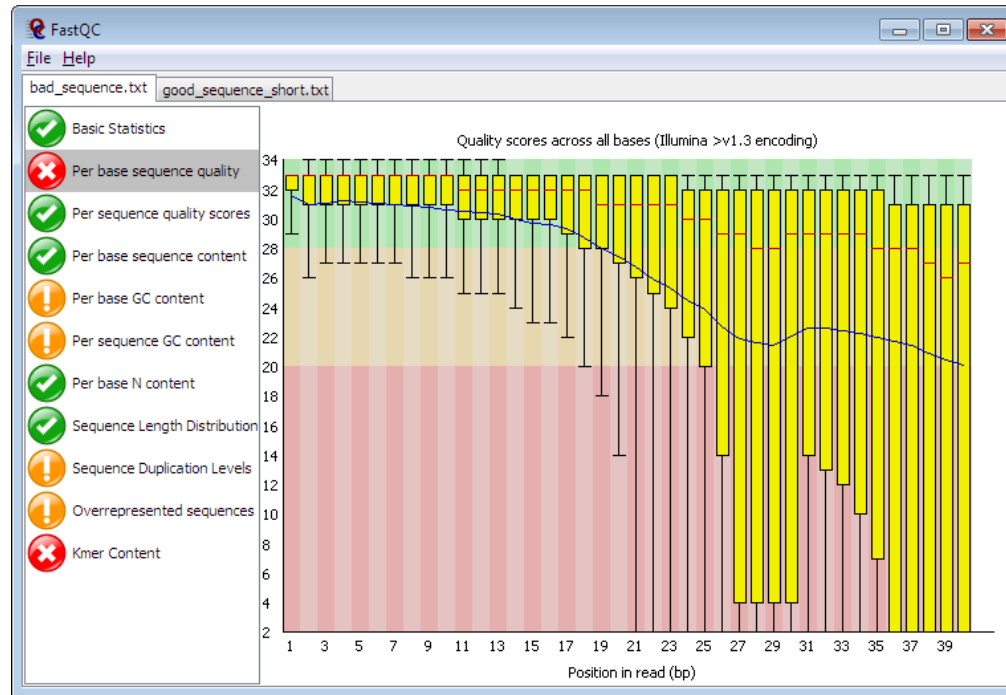
- <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

- FASTX

- http://hannonlab.cshl.edu/fastx_toolkit/commandline.html#fastx_barcode_splitter_usage

FastQC

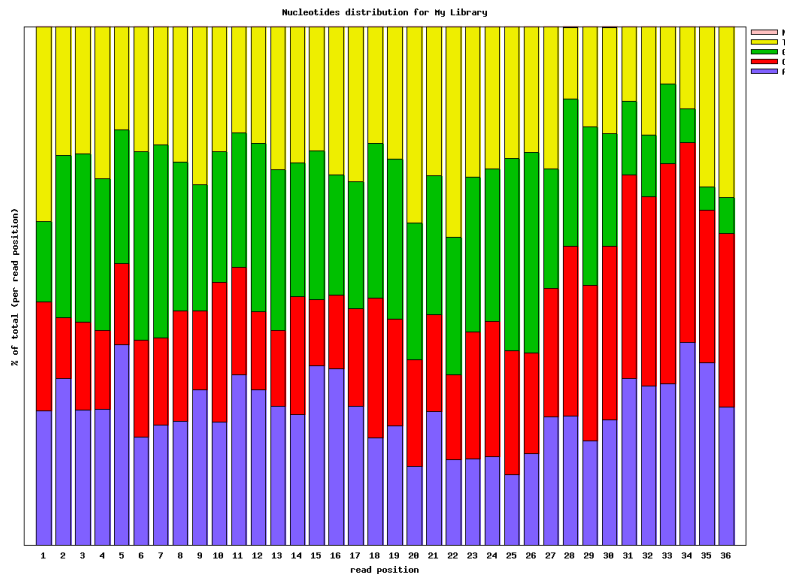
- 1枚のHTMLに複数の結果が、まとめられ出力される



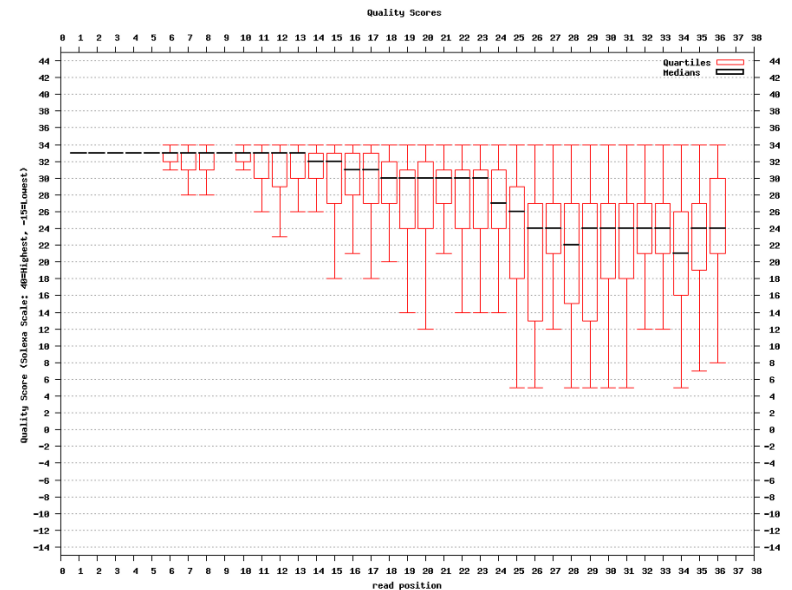
豆知識： `--nogroup` オプションで実行すると、1ベース毎の結果が表示される

FASTX

- 各項目ごとに、解析を行う
- Galaxyに入っている場合が多い



A relatively good quality library (median quality degrades towards later cycles):



豆知識 : CASAVA1.8以降では -Q33 オプションで実行する。

ときどきある質問

- Indexやバーコードなどの、特徴的な配列のサイクルのクオリティーが、下がる事がある。
 - Illumina シーケンサーは、同じサイクルで、同じ塩基を多数読むと、エラー率が高くなる。

RNA-Seq 解析について

- アダプター Trimming (必要なら)
- rRNA filtering
- マッピング
- 定量化
- 比較解析
- (De novo assembly)

rRNA filtering について

- ライブラリー作成時に、取り除けなかった rRNA のリードを除去する。
 - rRNA 配列に対して、Mapping を行い、Unmapped のリードを取りだす。
 - `samtools view -f 4 *.bam`

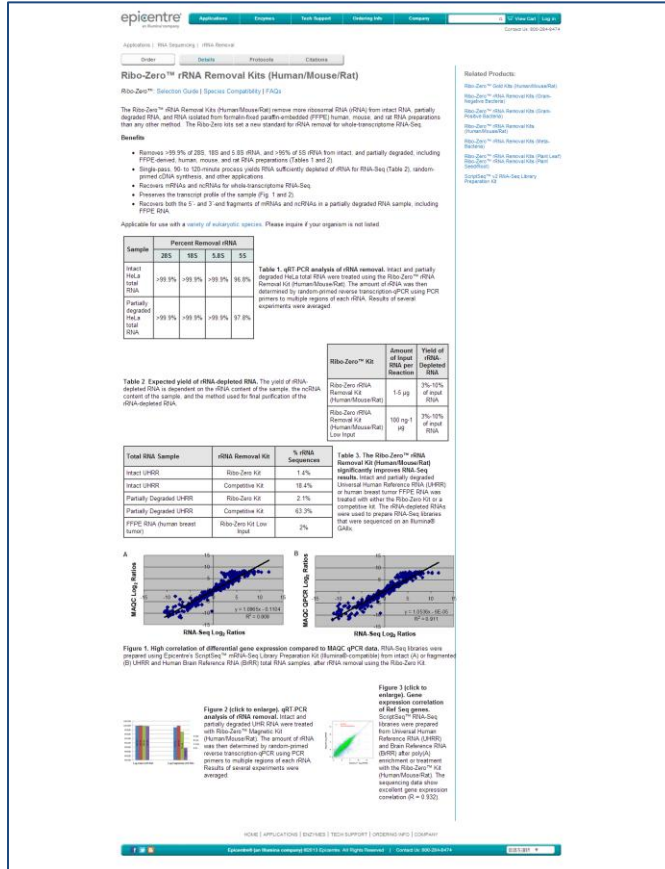
Mouse rRNA Reference : BK000964.1

<http://www.ncbi.nlm.nih.gov/nucore/BK000964>

Human rRNA Reference : U13369.1

<http://www.ncbi.nlm.nih.gov/nucore/U13369.1>

rRNA removal library kit



epigentrix Applications: RNA Sequencing | rRNA Removal

Order Details Protocols Catalogue

Ribo-Zero™ rRNA Removal Kits (Human/Mouse/Rat)

Ribo-Zero™ Selection Guide | Special Compatibility | FAQs

The Ribo-Zero™ rRNA Removal Kit (Human/Mouse/Rat) remove rRNA (ribosomal RNA) from total RNA, partially degraded rRNA, and rRNA isolated from formalin-fixed paraffin-embedded (FFPE) human, mouse, and rat RNA preparations from any other method. The Ribo-Zero kits are a new standard for rRNA removal for whole-transcriptome RNA-Seq.

Benefits

- Removes >99.9% of 28S, 18S and 5.8S rRNA, and >95% of 16S rRNA from intact and partially degraded, including FFPE-derived, human, mouse, and rat RNA preparations (Table 1 and 2).
- Single pass, 90- to 120-minute process yields RNA sufficiently depleted of rRNA for RNA-Seq (Table 2), undominated CDNA systems, and other applications.
- Recovers mRNAs and ncRNAs for whole-transcriptome RNA-Seq.
- Preserves the transcript profile of the sample (Fig. 1 and 2).
- Rescues both the 5' and 3' end fragments of mRNAs and ncRNAs in a partially degraded RNA sample including FFPE RNA.

Applicable for use with a variety of next-generation sequencers. Please inquire if your organism is not listed.

Sample	Percent Removal (rRNA)			SD
	28S	18S	5.8S	
Intact total RNA	>99.9%	>99.9%	>99.9%	18.8%
Partially degraded total RNA	>99.9%	>99.9%	>99.9%	17.8%

Table 1. qRT-PCR analysis of rRNA removal. Intact and partially degraded total RNA was treated using the Ribo-Zero™ rRNA Removal Kit (Human/Mouse/Rat). The amount of rRNA was then determined by random-primed reverse transcription-PCR using PCR primers to multiple regions of each rRNA. Results of several experiments were averaged.

Ribo-Zero™ Kit	Amount of Input RNA per Reaction	Yield of rRNA-Depleted RNA
Ribo-Zero rRNA Removal Kit (Human/Mouse/Rat)	1.0 µg	76%-10% of input RNA
Ribo-Zero rRNA Removal Kit (Human/Mouse/Rat) Low Input	100 pg 1 µg	75%-10% of input RNA

Table 2. Expected yield of rRNA-depleted RNA. The yield of rRNA-depleted RNA is dependent on the rRNA content of the sample, the rRNA content of the input, and the method used for final purification of the rRNA-depleted RNA.

Total RNA Sample	rRNA Removal Kit	% rRNA Sequenced
Intact LHRF	Ribo-Zero Kit	1.4%
Intact LHRF	Competitor Kit	18.4%
Partially Degraded LHRF	Ribo-Zero Kit	2.1%
Partially Degraded LHRF	Competitor Kit	63.9%
FFPE RNA (Human breast tumor)	Ribo-Zero Low Input	3%

Table 3. The Ribo-Zero™ rRNA Removal Kit (Human/Mouse/Rat) significantly improves RNA-Seq results. Intact and partially degraded Universal Human Reference RNA (LHRF) or human breast tumor FFPE RNA were treated with either the Ribo-Zero Kit or a competitor kit. The ribosomal rRNA was sequenced to measure the results that were sequenced on an Illumina HiSeq.

Figure 1. High correlation of differential gene expression compared to MACS qPCR data. RNA-Seq libraries were prepared using (A) Ribo-Zero™ rRNA Removal Kit (Human/Mouse/Rat) (LHRF) or Human Brain Reference RNA (LHRF) total RNA samples, after rRNA removal using the Ribo-Zero Kit.

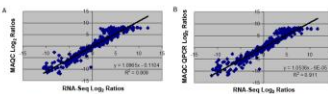


Figure 2. Link to enlarge: qRT-PCR analysis of rRNA removal. Intact and partially degraded LHRF RNA were treated with Ribo-Zero™ (Human/Mouse/Rat) or Human Brain Reference RNA (LHRF) total RNA samples, after rRNA removal using the Ribo-Zero™ (Human/Mouse/Rat). Results of several experiments were averaged.


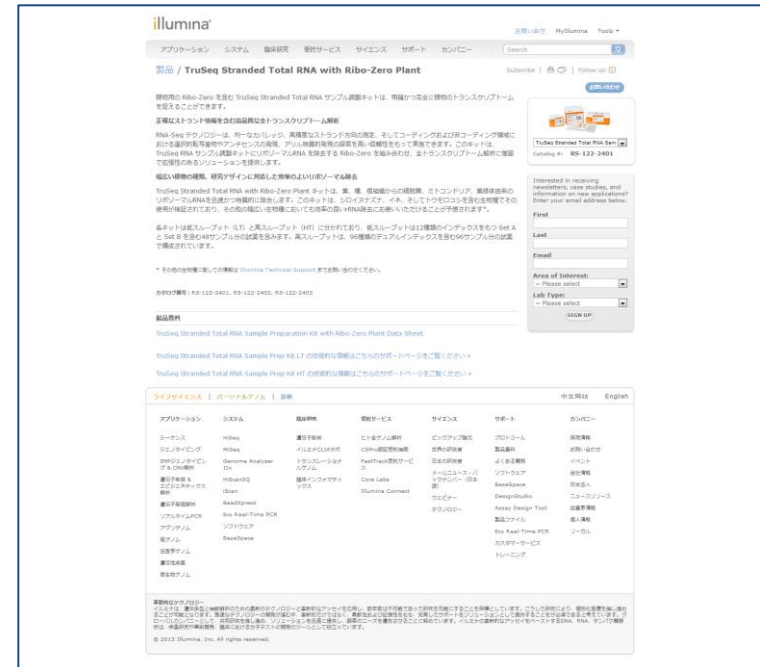



Figure 3. Link to enlarge: Gene expression correlation of Ribo-Zero treated RNA. Ribo-Zero™ rRNA Removal Kits were prepared from Universal Human Reference RNA (LHRF) and Human Brain Reference RNA (LHRF) total RNA samples, after rRNA removal using the Ribo-Zero™ (Human/Mouse/Rat). The sequencing data illustrate gene expression correlation (R = 0.82).



illumina 製品 / TruSeq Stranded Total RNA with Ribo-Zero Plant

詳細情報 Ribo-Zero を含む TruSeq Stranded Total RNA サンプル調製キットは、植物材料から抽出されたトランスクリプトームを改善します。

高品質のトランスクリプトームを確保するために、植物材料から抽出されたトランスクリプトームを改善します。

RNA-Seq サンプルは、均一なカバレッジ、高純度のストランド特異性を確保し、ココーディングおよび非コーディング領域にわたる最新の発見をアンチセンスの検出、アミノ末端側の延長を見出し、従来よりも改善されています。このキットは、TruSeq RNA サンプル調製キットと対応する rRNA を除去する Ribo-Zero キットを組み合わせた、ストランド特異的な植物で設計されたソリューションを提供します。

幅広い植物の種類、種多量に抽出された植物の rRNA 除去キット

TruSeq Stranded Total RNA with Ribo-Zero Plant キットは、高、純、高純度のストランド特異性を確保し、ココーディングおよび非コーディング領域にわたる最新の発見をアンチセンスの検出、アミノ末端側の延長を見出し、従来よりも改善されています。このキットは、TruSeq RNA サンプル調製キットと対応する rRNA 除去する Ribo-Zero キットを組み合わせた、ストランド特異的な植物で設計されたソリューションを提供します。

キットは高スループット (LIT) と低スループット (HT) に対応しており、高スループットには 12 種類のインデックスを含むセット A とセット B を含む 96 プラムのセットを提供します。高スループットは、96 種類のデュアルインデックスを含む 96 プラムのセットで提供されています。

* 本製品は生体試料での使用は Illumina Technical Support までお問い合わせください。

お問い合わせ先: 1-800-244-1444, 1-800-244-0400, 1-800-222-9400

製品情報

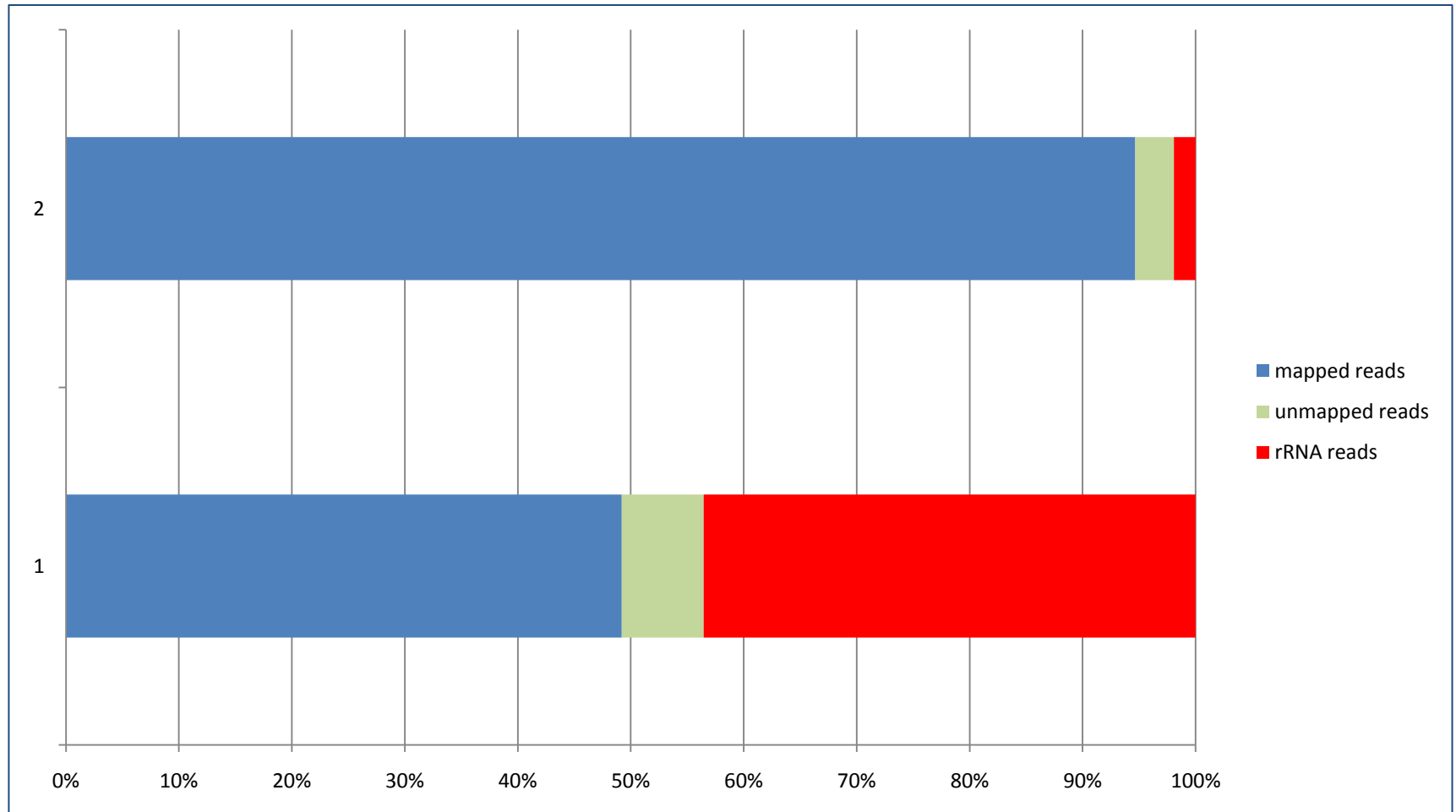
TruSeq Stranded Total RNA Sample Preparation Kit with Ribo-Zero Plant Data Sheet

TruSeq Stranded Total RNA Sample Prep Kit (LIT) の詳細な情報はこちらをクリックしてください >

TruSeq Stranded Total RNA Sample Prep Kit (HT) の詳細な情報はこちらをクリックしてください >

プラットフォーム	システム	読み取り	実行サービス	サイズ	サポート	オプション
シーラ	HiSeq	高通量	ヒトゲノム研究	ヒトゲノム研究	ブロード	研究
ジネティック	HiSeq	イルミナバ	ヒトゲノム研究	ヒトゲノム研究	ブロード	研究
マイクロアレイ	Genome Analyzer	ヒトゲノム研究	ヒトゲノム研究	ヒトゲノム研究	ブロード	研究
マイクロアレイ	DR	ヒトゲノム研究	ヒトゲノム研究	ヒトゲノム研究	ブロード	研究
マイクロアレイ	HiScanQ	ヒトゲノム研究	ヒトゲノム研究	ヒトゲノム研究	ブロード	研究
マイクロアレイ	Quant	ヒトゲノム研究	ヒトゲノム研究	ヒトゲノム研究	ブロード	研究
マイクロアレイ	NextFlow	ヒトゲノム研究	ヒトゲノム研究	ヒトゲノム研究	ブロード	研究
マイクロアレイ	Ion S5	ヒトゲノム研究	ヒトゲノム研究	ヒトゲノム研究	ブロード	研究
マイクロアレイ	Ion S5XL	ヒトゲノム研究	ヒトゲノム研究	ヒトゲノム研究	ブロード	研究
マイクロアレイ	Ion Torrent	ヒトゲノム研究	ヒトゲノム研究	ヒトゲノム研究	ブロード	研究

ライブラリー作成時にrRNAが良く取り除けた例と、悪い結果の例



悪い例の結果が出た場合

- ライブラリー作成、マニュアル、プロトコルを見直す
- ライブラリー再作成、再シーケンス

アダプター Trimming

- ソフトウェア

- FASTX “fastx_clipper”

- http://hannonlab.cshl.edu/fastx_toolkit/commandline.html#fastx_clipper_usage

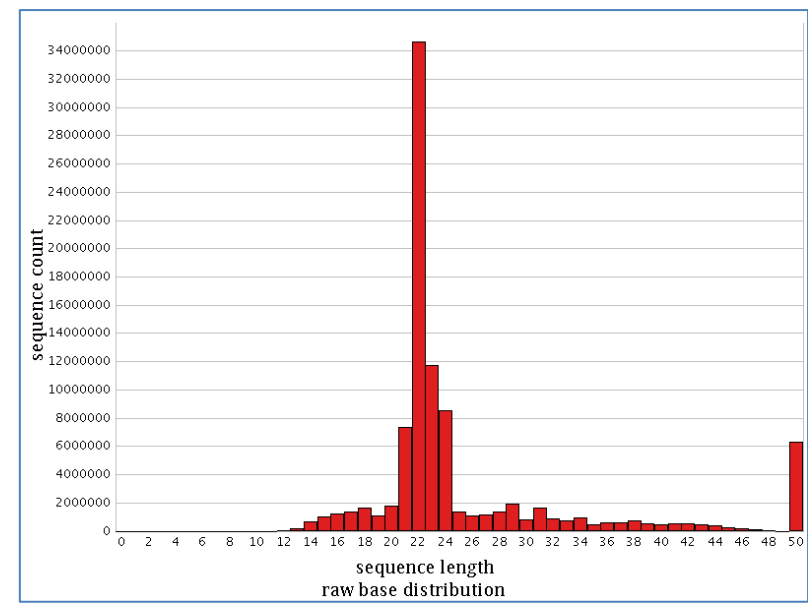
豆知識 : CASAVA1.8以降では `-Q33` オプションで実行する。

- *Cutadapt*

- <http://code.google.com/p/cutadapt/>

アダプター Trimming (例)

```
as01@lsa-serv-~/LS1952/fastqc
TCAGTGCACACAGAACTTTGTTGGAAATCTCGGGTGCCAGGAACCTCCN
+
@@0?0DDHFDGEBGGHGEIAHEECHEIGIIIIIFHGIIIIIIIIII#
@HWI-ST1394:57:C2GWJACXX:1:1101:3829:2117 1:N:0:CTTGTA
TATTAAGGTTCTGTTGTAATGGAAATCTCGGGTGCCAGGAACCTCCN
+
@CCFFEFFH7DFHFGGHIJJJJJEGJJJIGHDGHJI=D0FFIGGGHJJ#
@HWI-ST1394:57:C2GWJACXX:1:1101:3861:2119 1:N:0:CTTGTA
CTGTGCGTGTGACAGGGCTGATGGAAATCTCGGGTGCCAGGAACCTCCN
+
CCCCFFFFHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ#
@HWI-ST1394:57:C2GWJACXX:1:1101:3980:2128 1:N:0:CTTGTA
TCAGTGCACACAGAACTTTGTTGGAAATCTCGGGTGCCAGGAACCTCCN
+
CCCCFFFFHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ#
@HWI-ST1394:57:C2GWJACXX:1:1101:3787:2128 1:N:0:CTTGTA
TCAGTGCACACAGAACTTTGTAATGGAAATCTCGGGTGCCAGGAACCTCCN
+
CCCCFFFFHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ#
@HWI-ST1394:57:C2GWJACXX:1:1101:3936:2130 1:N:0:CTTGTA
CCCCCGCGGACCATGGAATCTCGGGTGCCAGGAACCTCAGTCACTGNN
+
CCCCFFFFHGHJJJJGHIJJJJJBGHI8=FHFHIIJJJJHHHHHH#
@HWI-ST1394:57:C2GWJACXX:1:1101:3843:2131 1:N:0:CTTGTA
AAGCTGCCAGTTGAAGAAGCTGTTGGAAATCTCGGGTGCCAGGAACCTCCN
+
CCCCFFFFHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ#
@HWI-ST1394:57:C2GWJACXX:1:1101:3921:2137 1:N:0:CTTGTA
ACAAGTCAGGCTCTTGGGACCTATGGAAATCTCGGGTGCCAGGAACCTCCN
+
CCCCFFFFHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ#
@HWI-ST1394:57:C2GWJACXX:1:1101:3831:2156 1:N:0:CTTGTA
ATGCACTGGTGAATCACTGTTGGAAATCTCGGGTGCCAGGAACCTCCN
+
CCBFFFFHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ#
@HWI-ST1394:57:C2GWJACXX:1:1101:3755:2157 1:N:0:CTTGTA
TCAGTGCACACAGAACTTTGTTGGAAATCTCGGGTGCCAGGAACCTCCN
+
CCCCFFFFHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ#
@HWI-ST1394:57:C2GWJACXX:1:1101:3884:2158 1:N:0:CTTGTA
ACTTGGGCCCCGGGTTCCCTCCGGGGCTACGGCTCTGACGCTGCTTN
+
CCCCFFFFHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ#
@HWI-ST1394:57:C2GWJACXX:1:1101:3863:2170 1:N:0:CTTGTA
TATTGCACCTGTCCGGGCTGTTGGAAATCTCGGGTGCCAGGAACCTCCN
+
CCCCFFFFHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ#
@HWI-ST1394:57:C2GWJACXX:1:1101:3966:2171 1:N:0:CTTGTA
CTTCTGATCGATGTGTCAGCTGCTGTTGGAAATCTCGGGTGCCAGGN
+
CCCCFFFFHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ#
@HWI-ST1394:57:C2GWJACXX:1:1101:3979:2189 1:N:0:CTTGTA
AAAGGTTGTTTGTAAAAATGGAAATCTCGGGTGCCAGGAACCTCAGTN
+
@CCFFEFFHHHHGHIJJJJJGEHIIIFGIJJDFHHCIIIIIIII#
@HWI-ST1394:57:C2GWJACXX:1:1101:3971:2204 1:N:0:CTTGTA
```



Trimming 前の,"50サイクル" Fastqデータ

Trimming 後のFastqデータの Length Distribution

Mapping for RNA-Seq

- TopHat2

– <http://tophat.cbcb.umd.edu/>

TopHat
A spliced read mapper for RNA-Seq

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

TopHat is a collaborative effort between the Center for Computational Biology at Johns Hopkins University, the Departments of Mathematics and Molecular and Cell Biology at the University of California, Berkeley, and the Department of Stem Cell and Regenerative Biology at Harvard University.

TopHat 2.0.9 release 6/28/2013
Note (6/29/2013): this version is slightly updated to handle 1-bp exons when using --OTF option.
Version 2.0.9 is a maintenance release providing better management of the transcriptome data files and fixes a few problems found in earlier releases:

- Solved parsing issues with some GFF3 files that could produce a crash with previous versions.
- Starting with this version TopHat2 will automatically check for consistency and, if needed, rebuild the existing transcriptome data files after critical updates of the GFF parser or a detected change of the underlying annotation data (GFF file).
- The output file `unmapped.bam` no longer contains multi-mapped reads (reads with too many alignments found), but only reads for which a suitable alignment could not be found under the current alignment constraints.
- A new output file: `sra_mv.txt` is now generated in the output directory, containing read (pair) input and mapping counts.
- Fixed a bug that added an extra XS tag in the output BAM file.
- Fixed a reporting bug that caused paired reads with a read containing its mate to be reported as unpaired.
- Fixed a bug in `bam2fastx` utility that caused the `-M` (mapped-only) option to be ignored. (Note: this option is not used within TopHat).
- In `tophat-fusion-post`: fixed a bug that caused two genes of a fusion gene to sometimes be incorrectly ordered and reported.

TopHat2 paper published 4/25/2013
The TopHat2 paper has been published in *Genome Biology*.
• This simulation data set (error-free) is available here.

TopHat 2.0.8 release 2/26/2013
Note (4/12/2013): patched version 2.0.8b was released in order to provide compatibility with Bowtie v1.0.0
Version 2.0.8 is a quick fix release addressing the following issues:

- This version correctly handles the newest version of Bowtie2 v2.1.0.
- The segment mapping slow-down introduced by some Bowtie2 parameter changes in version 2.0.7 is now corrected.

TopHat 2.0.7 release 1/23/2013

Site Map
Home
Getting started
Manual
Index and annotation downloads
FAQ
Protocol

News and updates
New releases and related tools will be announced through the Bowtie [mailing list](#).

Get Help
Questions and comments about TopHat can be posted on the [TopHat Users Google Group](#). Please use tophat.cufirika@gmail.com for private communications only. Please do not email technical questions to TopHat contributors directly.

Releases
version 2.0.9 6/28/2013

Kim et al. *Genome Biology* 2013, **14**:R66
<http://genomebiology.com/2013/14/R66>

METHOD Open Access

TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions

Daehwan Kim^{1,2,3*}, Geo Pertea³, Cole Trapnell^{3,6}, Harold Pimentel⁷, Ryan Kelley⁸ and Steven L Salzberg^{1,4}

Abstract
TopHat is a popular spliced aligner for RNA-sequence (RNA-seq) experiments. In this paper, we describe TopHat2, which incorporates many significant enhancements to TopHat. TopHat2 can align reads of various lengths produced by the latest sequencing technologies, while allowing for variable-length indels with respect to the reference genome. In addition to *de novo* spliced alignment, TopHat2 can align reads across fusion breaks, which can occur after genomic translocations. TopHat2 combines the ability to identify novel splice sites with direct mapping to known transcripts, producing sensitive and accurate alignments, even for highly repetitive genomes or in the presence of pseudogenes. TopHat2 is available at <http://ccb.jhu.edu/software/tophat>.

Background
RNA-sequencing technologies [1], which sequence the RNA molecules being transcribed in cells, allow exploration of the process of transcription in exquisite detail. One of the primary goals of RNA-sequencing analysis software is to reconstruct the full set of transcripts (isoforms) of genes that were present in the original cells. In addition to the transcript structures, experimenters need to estimate the expression levels for all transcripts. The first step in the analysis process is to map the RNA-sequence (RNA-seq) reads against the reference genome, which provides the location from which the reads originated. In contrast to DNA-sequence alignment, RNA-seq mapping algorithms have two additional challenges. First, because genes in eukaryotic genomes contain introns, and because reads sequenced from mature mRNA transcripts do not include these introns, any RNA-seq alignment program must be able to handle gapped (or spliced) alignment with very large gaps. In mammalian genomes, introns span a very wide range of lengths, typically from 50 to 100,000 bases, which the alignment algorithm must accommodate. Second, the presence of processed pseudogenes, from which some or all introns have been removed, may cause many exon-spanning reads to map incorrectly. This is particularly acute for the human genome, which contains over 14,000 pseudogenes [2].

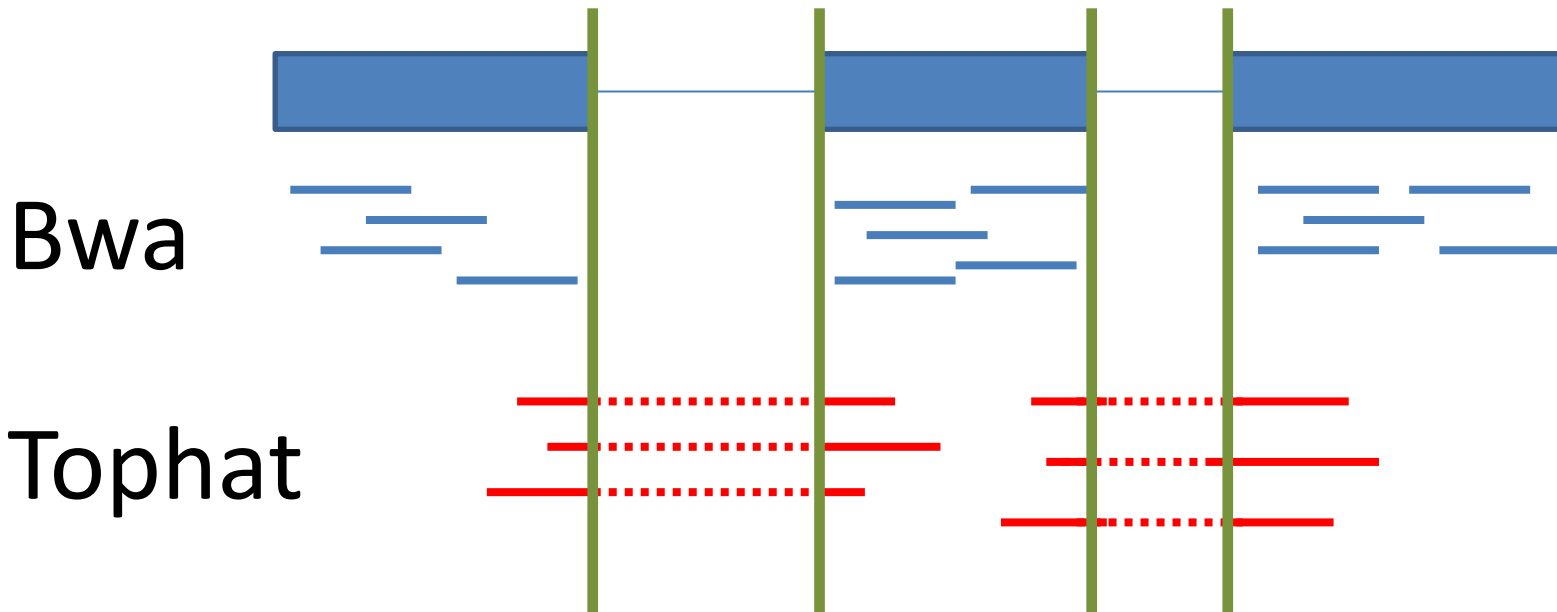
In the most recent Ensembl GRCh37 gene annotations, the average length of a mature mRNA transcript in the human genome is 2,227 bp long and the average exon length is 235 bp. The average number of exons per transcript is 9.5. Assuming that sequencing reads are uniformly distributed along a transcript [3], we would expect 33 to 38% of 100 bp reads from an RNA-seq experiment to span two or more exons. Note that this proportion increases significantly as read length increases from 50 to 150 bp (see Additional file 1 for more details).

More important for the alignment problem is that around 20% of junction-spanning reads extend by 10 bp or less into one of the exons they span. These small 'anchors' make it extremely difficult for alignment software to map reads accurately, particularly if the algorithm relies (as most do) on an initial mapping of fixed-length k-mers to the genome. This initial mapping, using exact matches of k-mers, is crucial for narrowing down the search space into small local regions in which a read is likely to align. If a read extends only a few bases into one of two adjacent exons, then it often happens that the read will align equally well, but incorrectly, with the sequence of the intervening intron. For example, as illustrated in Figure 1, suppose that read r spans exons e_1 and e_2 , extending only four bases into e_2 . Suppose also that that e_2 extending only

* Correspondence: inf@tophat.cbcb.umd.edu
¹Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD, 20742, USA
Full list of author information is available at the end of the article

BioMed Central ©2013 Kim et al.; licensee BioMed Central Ltd. This is an open access article distributed under the terms of the Creative Commons Attribution License

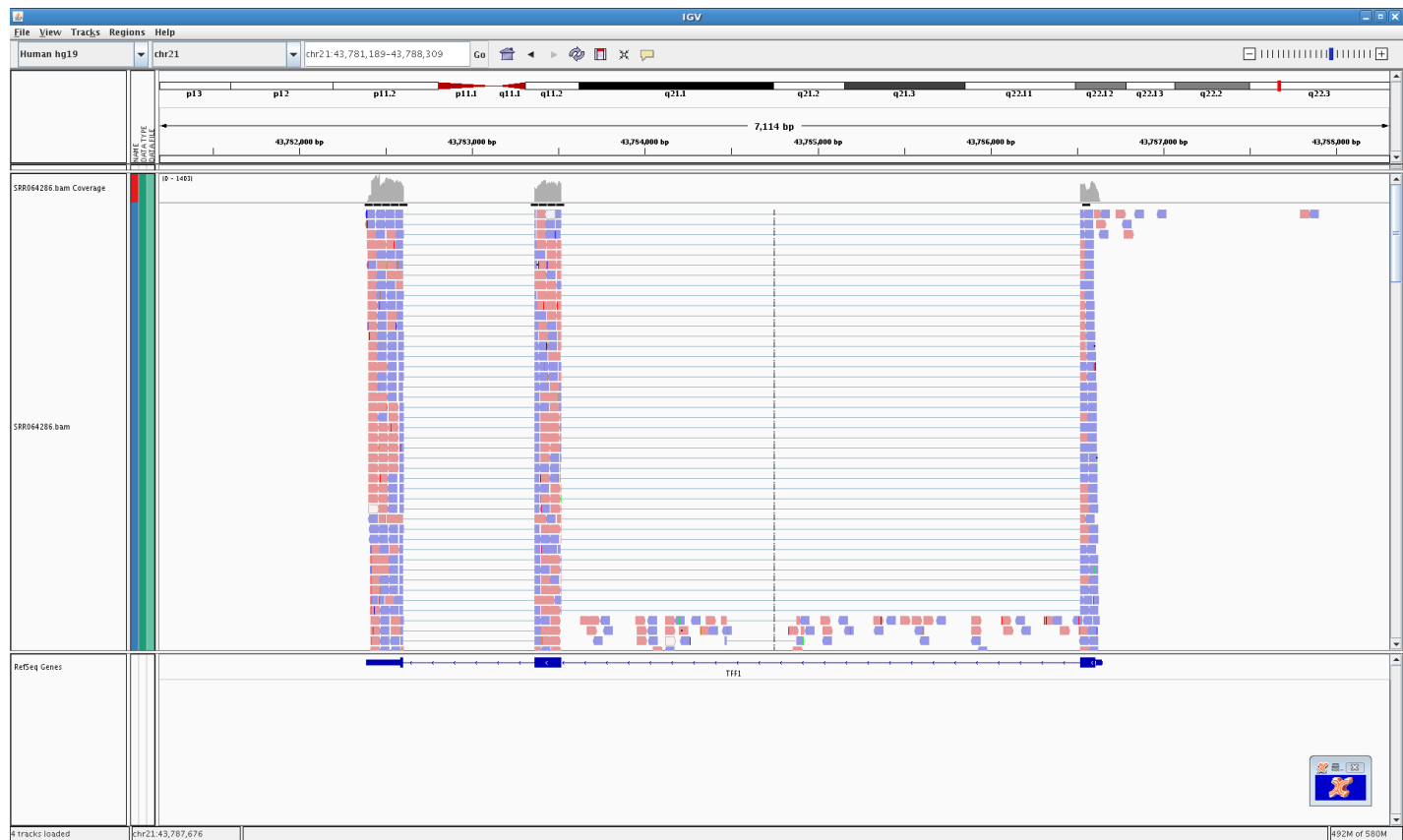
Mapping の違い



Bwa

Tophat

Tophat による mapping (例)



定量 (マップされたリードの数から、normalizeして値を算出)



Gene	Mapped Read count	exon size	Total mapped read count	RPM (read per Million)	RPKM (RPM per kilo exon)
A	400	1000	10,000,000	40 (=400*1,000,000/10,000,000)	<u>40</u>
B	200	500		20	<u>40</u>
C	200	400		20	<u>50</u>
...

cufflinks

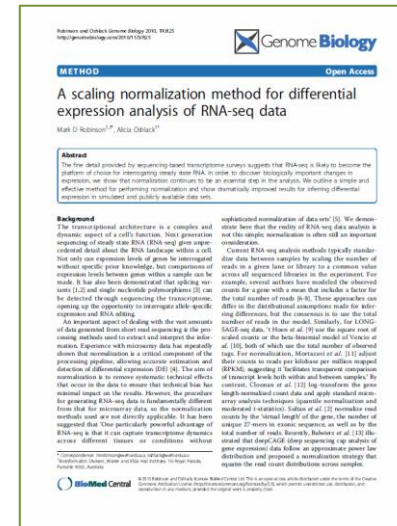
- Cufflinks

- <http://cufflinks.cbcb.umd.edu/>

- アノテーション情報とマッピング結果より、FPKM を算出。

定量、比較の問題

- RPKM(FPKM)は、遺伝子(exon size)の大きさや、高発現遺伝子の影響により、結果がばらつく。
 - TMM (Trimmed Mean of M-values) による正規化



Rによる比較解析

- DESeq
- edgeR
- ...

The screenshot shows the Bioconductor website homepage. The header features the Bioconductor logo and navigation links: Home, Install, Help, Developers, and About. A search bar is located in the top right corner. The main content area is divided into several sections:

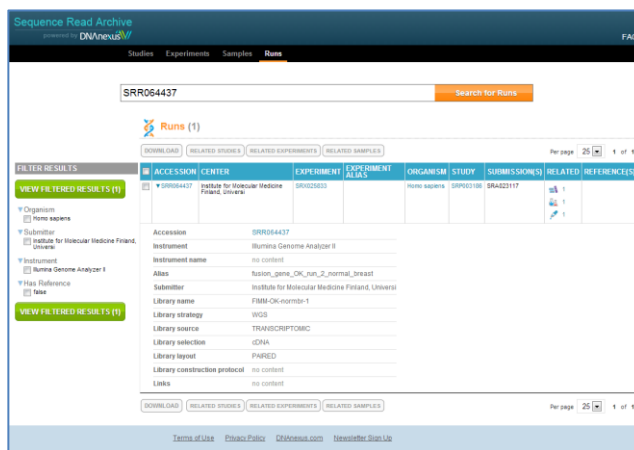
- About Bioconductor:** A section describing the project's goals and resources, including links to software packages and an Amazon Machine Image (AMI).
- Use Bioconductor for...:** A section listing various applications such as Sequence Data, Annotation, Transcription Factors, and Counting Reads for Differential Expression.
- Mailing Lists:** A section with a search bar and a list of recent email threads, including "Re: 450k annotation packages compatib...", "Re: autoplot transcriptDb error with ...", "Re: ANCOVA microarray time-course con...", and "Re: ANCOVA microarray time-course".
- Events:** A section listing upcoming events, such as "EMBO Practical Course on Analysis of High-Throughput Sequencing Data", "RNA-Seq analysis using Bioconductor", "EMBO Practical Course: Bioinformatics and statistics for large-scale data", "Bioconductor European Developers' Workshop", and "Next Generation Data Analysis".
- Tweets:** A section displaying recent tweets from the Bioconductor account, including announcements about package releases and workshops.

Rによる解析

- 良いところ
 - Normalize、正規化、比較解析まで、パッケージ化されている
 - 正規化される事により、バイアスの少ない結果が出る
- 少しめんどくさいところ
 - Rの使い方を覚える
 - BAMから、タグカウントの情報を作成する。
 - Samtools、HTSeqなどを使用する

RNA-Seq 解析例

登録データ	サンプル	Library - Sequence
SRR064437	正常ヒト胸腺由来cDNAのRNA-seqデータ	Non directional RNA-Seq Paired End Sequence
SRR064286	ヒトMCF-7 breast cancer cell line由来のRNA-seqデータ	Non directional RNA-Seq Paired End Sequence



Sequence Read Archive
powered by DNAnexus

Search for Runs: SRR064437

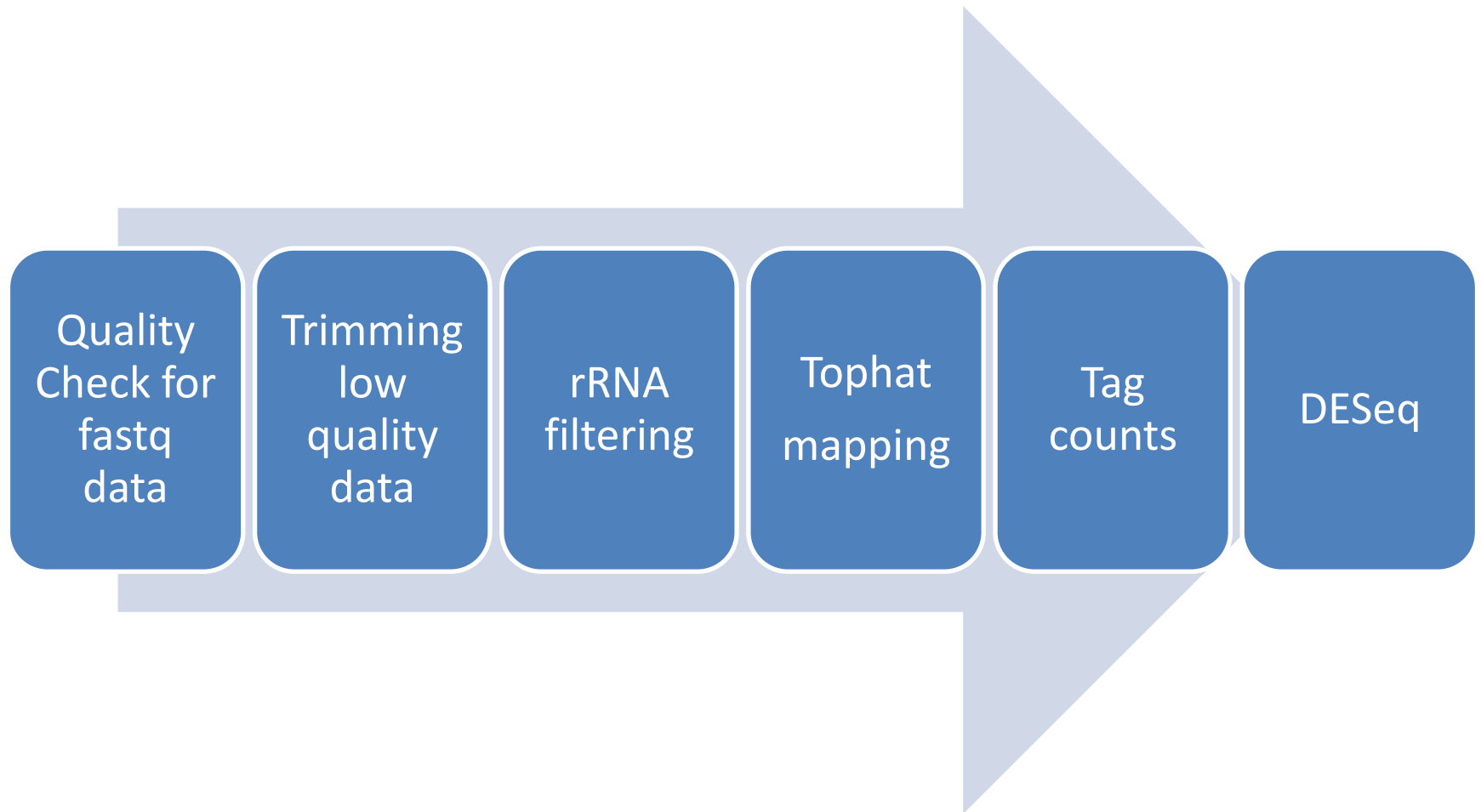
Runs (1)

ACCESSION	CENTER	EXPERIMENT	EXPERIMENT ALIAS	ORGANISM	STUDY	SUBMISSIONS	RELATED	REFERENCE(S)
SRR064437	Institute for Molecular Medicine Finland, University of Helsinki	SRR025853		Human sapiens	SRR02108	SRR023117		

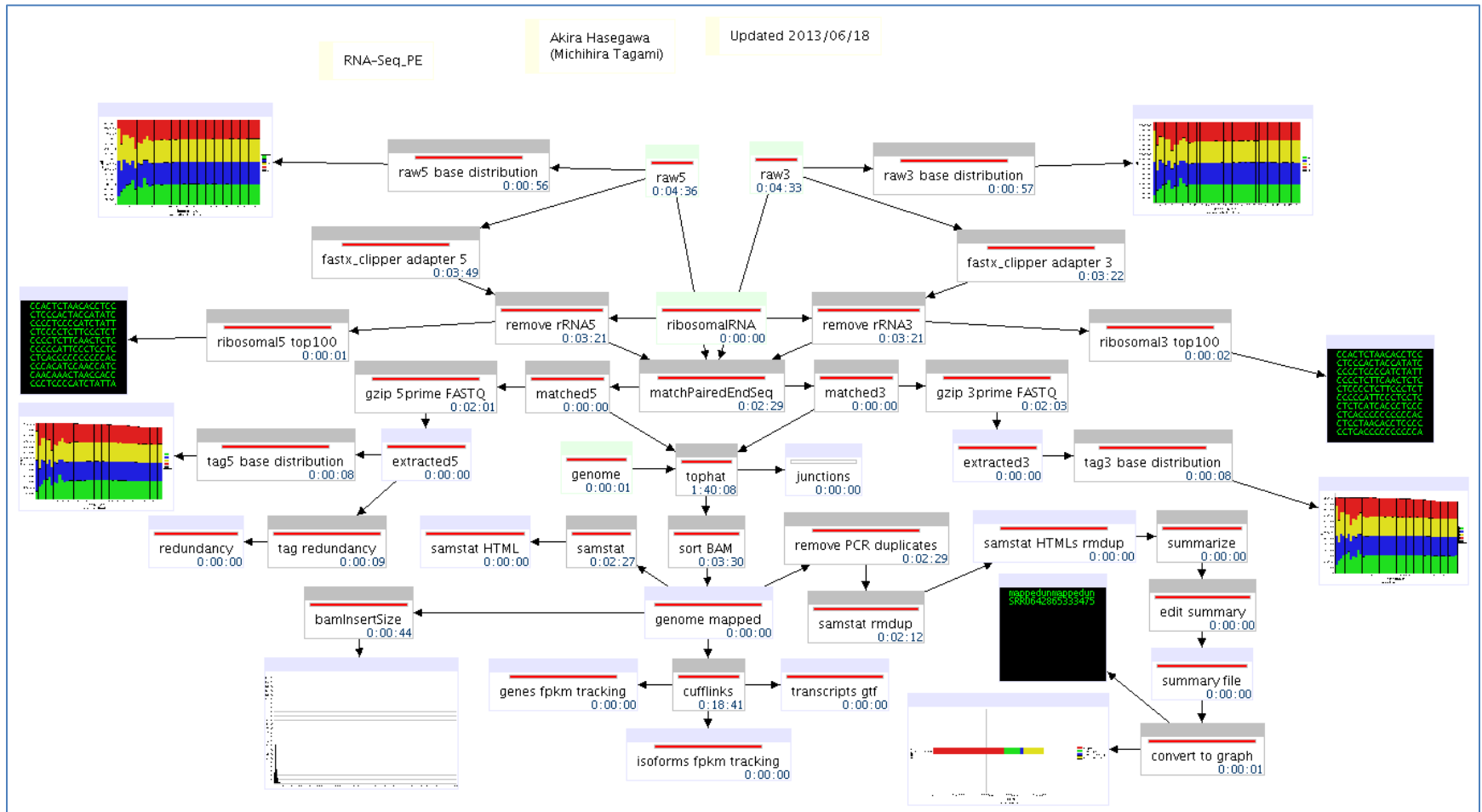
Accession: SRR064437
 Instrument: Illumina Genome Analyzer II
 Instrument name: no content
 Alias: rna_seq_demo_OK_nor_2_normal_breast
 Submitter: Institute for Molecular Medicine Finland, University of Helsinki
 Library name: FIMM-OK-norbid-1
 Library strategy: WGS
 Library source: TRANSCRIPTOMIC
 Library selection: cDNA
 Library layout: PAIRED
 Library construction protocol: no content

SRA : <http://www.ncbi.nlm.nih.gov/Traces/sra/>
 DRA : <http://trace.ddbj.nig.ac.jp/dra/index.html>

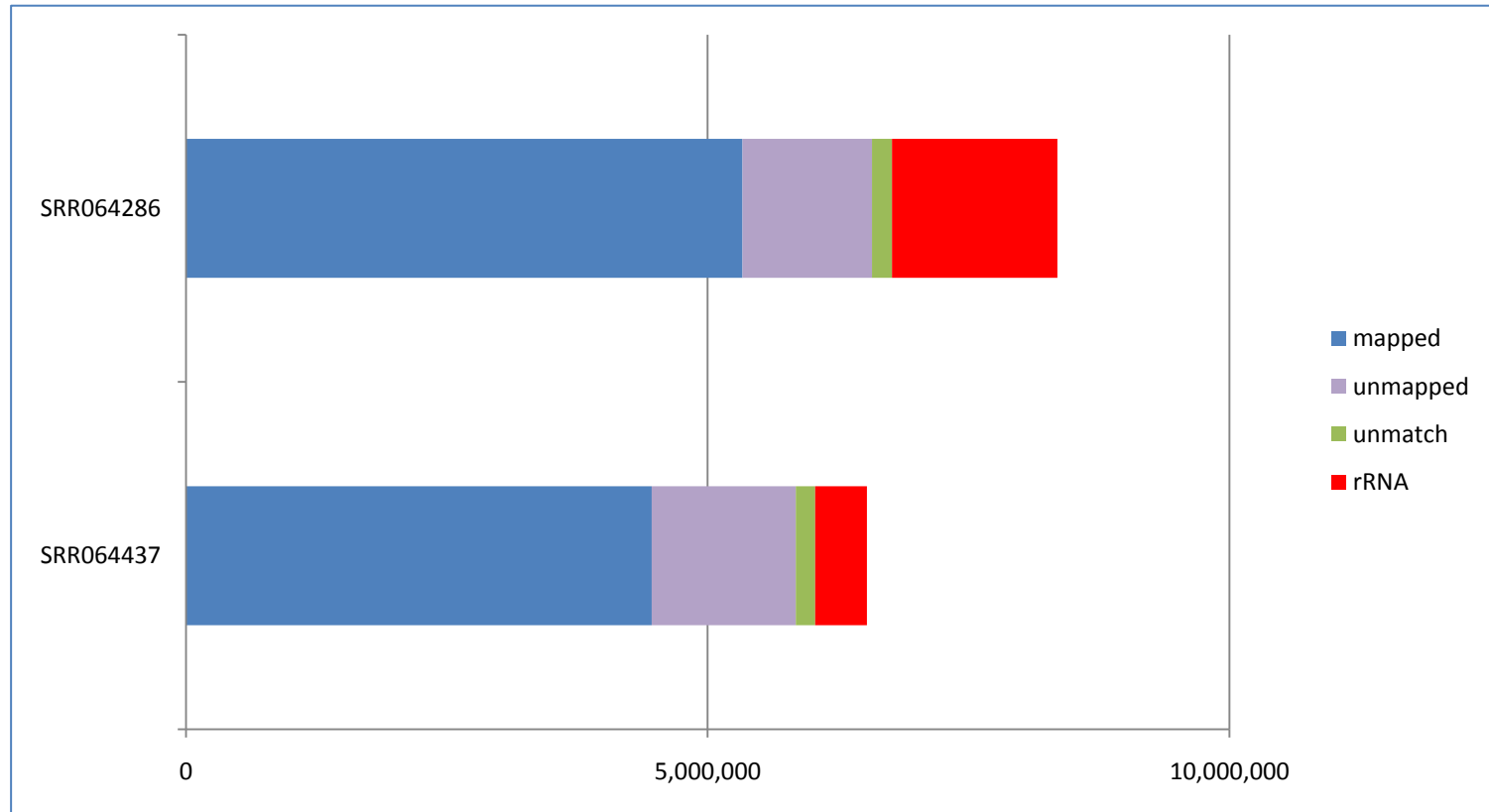
RNA-Seq 解析例 (workflow)



Mapping workflow with “moirai”



Summary for mapping



Tag count (HTSeq)

- Mapping 結果のBAMを samtools で、SAMファイルに変換
 - `samtools sort SRR064437.bam SRR064437_sorted`
 - `samtools view SRR064437_sorted.bam > SRR064437_sorted.sam`
 - HTSeqにより、タグカウント
 - `htseq-count SRR064437_sorted.sam gencode.v18.annotation.gtf > SRR064437_tag-count.txt`
-

- HTSeq
 - <http://www-huber.embl.de/users/anders/HTSeq/doc/index.html>
- 使用したアノテーションファイル
 - “gencode.v18.annotation.gtf”
 - <http://www.gencodegenes.org/>

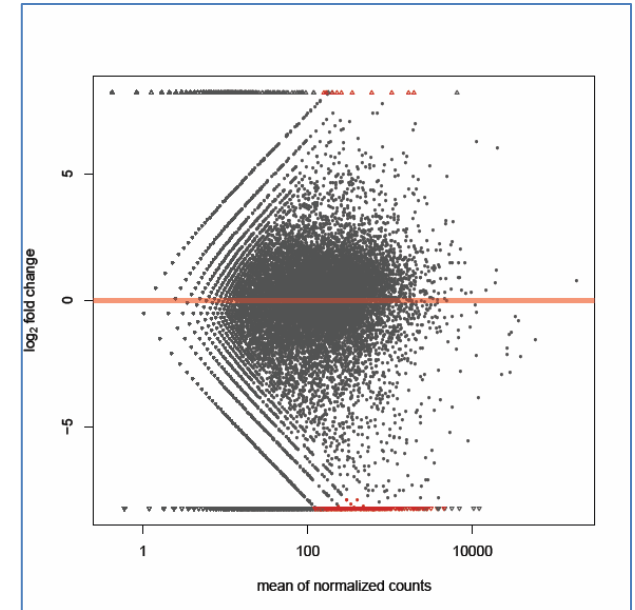
Tag countの結果 (HTSeq)

```
root@lsx013:~# cat > /dev/null
Ensembl ID      Normal hit tag count  MCF-7 tag count
NSG00000000003.10 250      70
NSG00000000005.5 28        0
NSG000000000419.8 55        190
NSG000000000457.8 62        63
NSG000000000460.12 18        72
NSG000000000938.8 180       0
NSG000000000971.11 387       0
NSG00000001036.8 110       421
NSG00000001084.6 91        182
NSG00000001167.10 78        168
NSG00000001460.13 39        28
NSG00000001461.12 110       134
NSG00000001497.12 157       356
NSG00000001561.6 75        27
NSG00000001617.7 291       279
NSG00000001626.9 0         0
NSG00000001629.5 148       513
NSG00000001630.11 11        73
NSG00000001631.10 117       166
NSG00000002016.12 46        37
NSG00000002079.8 2         2
NSG00000002330.9 128       150
NSG00000002549.8 282       299
NSG00000002586.13 0         0
NSG00000002587.5 18        0
NSG00000002726.15 4         0
NSG00000002745.8 0         0
NSG00000002746.9 0         0
NSG00000002822.11 78        194
NSG00000002834.13 936       1946
NSG00000002919.10 84        197
NSG00000002933.3 151       0
NSG00000003056.3 220       457
NSG00000003096.9 20        3
NSG00000003137.4 360       276
NSG00000003147.13 83        313
NSG00000003249.9 25        600
NSG00000003393.10 91        77
NSG00000003400.10 93        2
NSG00000003402.14 556       95
NSG00000003436.10 270       30
NSG00000003509.11 37        61
NSG00000003756.12 705       258
NSG00000003987.9 1         11
NSG00000003989.12 267       2606
NSG00000004059.6 250       937
NSG00000004139.9 112       135
NSG00000004142.7 280       712
NSG00000004399.8 900       520
```

- 1カラム目 : Ensembl ID
- 2カラム目 : 正常ヒトのタグカウント
- 3カラム目 : MCF-7 breast cancer cell line タグカウント

DESeq output

En ID	Gene Symbol	chromosome position	SRR064437 正常ヒト胸腺	SRR06428 6 breast cancer cell line	p.value	q.value	ranking	logratio
110	ENSG00000104332.7	SFRP1 chr8:41119480-41167016	1429	1	0.000187342	0.045786816	109	-10.98215486
111	ENSG00000211897.3	IGHG3 chr14:106235438-106237742	1284	1	0.000187724	0.045786816	110	-10.82779414
112	ENSG00000146122.12	DAAM2 chr6:39760141-39872648	470	0	0.000187857	0.045786816	111	#NAME?
113	ENSG00000186642.11	PDE2A chr11:172287184-72385635	1493	1	0.000188301	0.045786816	112	-11.0453631
114	ENSG00000184811.3	TUSC5 chr17:1182956-1204281	469	0	0.000188513	0.045786816	113	#NAME?
115	ENSG00000113140.6	SPARC chr5:15104056-151066726	3627	0	0.000188641	0.045786816	114	#NAME?
116	ENSG00000142910.10	TNFR1L chr1:32042115-32053288	1591	1	0.00019084	0.045917872	115	-11.13708278
117	ENSG0000010210.2	SEMA3C chr3:52467068-52470101	1735	1	0.000196405	0.04670816	116	-11.2620846
118	ENSG00000135046.9	ANKA1 chr9:756667-7578309	1071	1	0.000197555	0.04670816	117	-10.56610742
119	ENSG00000127954.8	STAP4 chr7:87905743-87936206	1040	1	0.000200308	0.04670816	118	-10.52373247
120	ENSG0000007237.13	GAS7 chr17:9813925-10101868	450	0	0.000202087	0.04670816	119	#NAME?
121	ENSG00000160182.2	TFF1 chr21:43782390-43786703	0	4000	0.00020527	0.04670816	120	Inf
122	ENSG00000169554.12	ZEB2 chr2:145141647-145282147	442	0	0.000208533	0.04670816	121	#NAME?
123	ENSG00000112233.8	LTF chr3:46477135-46526724	3872	0	0.000208588	0.04670816	122	#NAME?
126	ENSG0000011716.8	LDHB chr12:21788275-21910791	434	0	0.00021547	0.04670816	124	#NAME?
126	ENSG00000006518.12	COL17A1 chr10:105791043-105845760	431	0	0.000218207	0.04670816	125	#NAME?
127	ENSG00000004776.7	HSPB6 chr19:36245468-36248980	900	1	0.000219171	0.04670816	126	-10.31514585
128	ENSG00000104518.6	GSDMD chr8:144635376-144645232	428	0	0.000221022	0.04670816	127.5	#NAME?
129	ENSG00000188257.6	PLA2G2A chr1:20301924-20306932	428	0	0.000221022	0.04670816	127.5	#NAME?
130	ENSG000000091986.11	CDC80 chr3:112323406-112368377	888	1	0.000221408	0.04670816	129	-10.29578052
131	ENSG00000154783.6	FOD5 chr3:14860468-14975895	427	0	0.000221978	0.04670816	130	#NAME?
132	ENSG00000187955.7	COL14A1 chr8:121072018-121384275	425	0	0.000223917	0.04670816	131	#NAME?
133	ENSG00000120318.11	ARAP3 chr5:141032967-141061788	423	0	0.000225893	0.04670816	132	#NAME?
134	ENSG00000167588.8	GPD1 chr12:50497601-50505102	861	1	0.000226894	0.04670816	133	-10.25123408
135	ENSG00000123689.5	GOS2 chr1:209848764-209849733	417	0	0.000232053	0.04670816	134.5	#NAME?
136	ENSG00000171115.3	GIMAP8 chr7:150147171-150176480	417	0	0.000232053	0.04670816	134.5	#NAME?
137	ENSG00000136492.4	BRIP1 chr17:59758626-59940882	0	827	0.000233065	0.04670816	136	Inf
138	ENSG00000143248.8	RG55 chr1:163080910-163291577	828	1	0.000234551	0.04670816	137	-10.19485161
139	ENSG00000170801.5	HTRA3 chr4:8271491-8308838	414	0	0.000235268	0.04670816	138	#NAME?
140	ENSG00000149451.13	ADAM33 chr20:3648611-3662893	411	0	0.000238577	0.04670816	139	#NAME?
141	ENSG00000229124.2	VJM-AS1 chr10:17255237-17271984	4177	0	0.000239643	0.04670816	140	#NAME?
142	ENSG00000130592.9	LSP1 chr11:1874199-1913497	410	0	0.000239702	0.04670816	141.5	#NAME?
143	ENSG00000134853.7	PDGFRA chr4:55095263-55164414	410	0	0.000239702	0.04670816	141.5	#NAME?
144	ENSG00000127329.10	PTRFB chr12:70910629-71031220	801	1	0.000241708	0.046769726	143	-10.14702309
145	ENSG00000196542.4	SPTSS8 chr3:161062579-161090668	0	4642	0.000244855	0.047049656	144	Inf
146	ENSG00000152661.7	GJA1 chr6:121756837-121770873	400	0	0.000251578	0.047916772	145	#NAME?
147	ENSG00000154258.12	ABC9A chr17:66970628-67057205	399	0	0.000252832	0.047916772	146	#NAME?



Up-regulated
 TOP1 : DSCAM-AS1
 TOP2 : TFF1
 TOP3 : BRIP1
 ...

pValue でソートして、“正常ヒト”に対して、“breast cancer”で、Up-regulatedされた遺伝子

TFF1

Open

Oncogene (2011) 30, 3261–3273
© 2011 Macmillan Publishers Limited. All rights reserved 0950-9230/11
www.nature.com/onc



ORIGINAL ARTICLE

Deficiency in trefoil factor 1 (TFF1) increases tumorigenicity of human breast cancer cells and mammary tumor development in TFF1-knockout mice

E. Buache^{1,3}, N. Etique^{1,3}, F. Alpy¹, I. Stoll¹, M. Muckensturm², B. Reina-San-Martin¹, MP Chenard², C. Tomasetto¹ and MC Rio¹

¹Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), CNRS UMR 7104, INSERM U964, Université de Strasbourg, Illkirch, France and ²Service d'Anatomie Pathologique Générale, Centre Hospitalier Universitaire de Hautepierre, Strasbourg, France

Although trefoil factor 1 (TFF1; previously named pS2) is abnormally expressed in about 50% of human breast tumors, its physiopathological role in this disease has been poorly studied. Moreover, controversial data have been reported. TFF1 function in the mammary gland therefore needs to be clarified. In this study, using retroviral vectors, we performed TFF1 gain- or loss-of-function experiments in four human mammary epithelial cell lines: normal immortalized TFF1-negative MCF10A, malignant TFF1-negative MDA-MB-231 and malignant TFF1-positive MCF7 and ZR75.1. The expression of TFF1 stimulated the migration and invasion in the four cell lines. Forced TFF1 expression in MCF10A, MDA-MB-231 and MCF7 cells did not modify anchorage-dependent or -independent cell proliferation. By contrast, TFF1 knockdown in MCF7 enhanced softagar colony formation. This increased oncogenic potential of MCF7 cells in the absence of TFF1 was confirmed *in vivo* in nude mice. Moreover, chemically induced tumorigenesis in TFF1-deficient (TFF1-KO) mice led to higher tumor incidence in the mammary gland and larger tumor size compared with wild-type mice. Similarly, tumor development was increased in the TFF1-KO ovary and lung. Collectively, our results clearly show that TFF1 does not exhibit oncogenic properties, but rather reduces tumor development. This beneficial function of TFF1 is in agreement with many clinical studies reporting a better outcome for patients with TFF1-positive breast primary tumors.

Oncogene (2011) 30, 3261–3273; doi:10.1038/onc.2011.41; published online 28 February 2011

Keywords: TFF1/pS2; breast cancer; gain- and loss-of-function; human mammary cell lines; tumorigenicity; TFF1-KO mice

Correspondence: Dr MC Rio, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), CNRS UMR 7104, INSERM U964, Université de Strasbourg, Illkirch, France.
E-mail: rio@igbmc.fr

³Co-first authors.

Received 29 July 2010; revised 16 January 2011; accepted 20 January 2011; published online 28 February 2011

Introduction

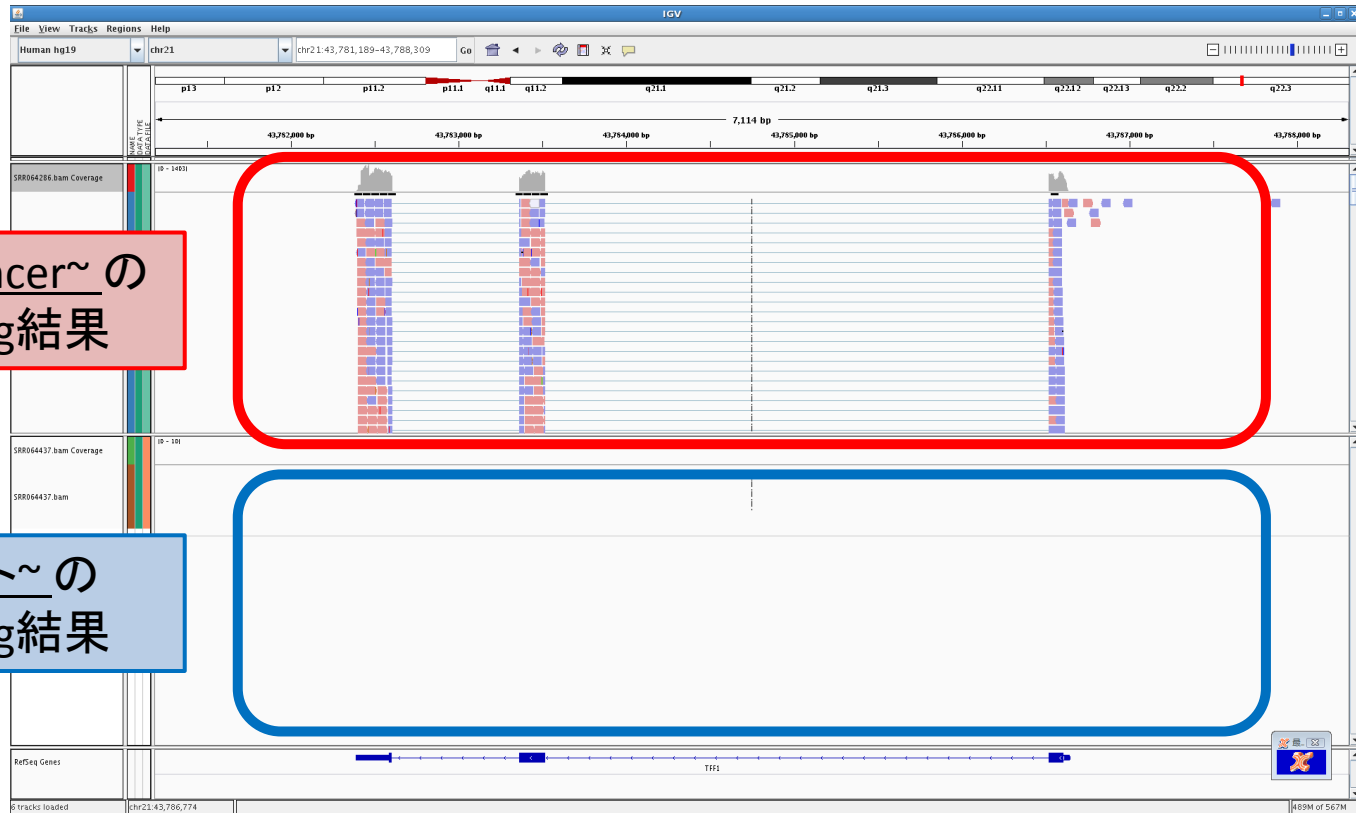
Trefoil factor 1 (TFF1; previously named pS2) (HUGO Gene Nomenclature Committee; <http://www.genenames.org>) is a small cysteine-rich acidic secreted protein (Thim, 1997; Ribieras *et al.*, 1998). It is constitutively and strongly expressed in the stomach, where it has a key role (Rio *et al.*, 1988). Indeed, TFF1 is essential for the normal differentiation of the gastric glands (Bossmeyer-Pourie *et al.*, 2002; Karam *et al.*, 2008). Moreover, by interacting with mucins, TFF1 participates in the correct organization of the mucus layer and in the gastric mucosa protection (Tomasetto *et al.*, 2000). Transgenic mice overexpressing TFF1 have an increased resistance to ulceration (Playford *et al.*, 1996). TFF1 is also expressed in the inflamed or damaged gastrointestinal tract, supporting the hypothesis that it mediates repair processes (Rio *et al.*, 1991; Kjellev, 2009). Indeed, TFF1 promotes epithelial restitution after injury and protects the integrity of the epithelial barrier (Hoffmann, 2005). Moreover, TFF1 is also expressed, but to a lesser extent, by normal epithelial cells of numerous organs (eyes, lung, ovary and salivary gland) (Regalo *et al.*, 2005; Madsen *et al.*, 2007; Buron *et al.*, 2008). To date, the TFF1 function during malignant processes is not clearly defined, as epithelial cell transformation might lead to downregulation of TFF1 expression (that is, in the stomach) or to the induction of TFF1 expression (that is, in various organs).

In the stomach, TFF1-deficient mice (TFF1-KO) develop antro-pyloric hyperplasia and dysplasia, leading to adenomas and intraepithelial or intramucosal carcinomas (Lefebvre *et al.*, 1996). Epithelial progenitors are amplified and are more invasive (Karam *et al.*, 2008). It has therefore been proposed that TFF1 functions as a gastric tumor suppressor gene. Strongly supporting this hypothesis, 50% of human gastric tumors are devoid of TFF1 because of deletions, mutations or methylation of the TFF1 gene (Ribieras *et al.*, 1998; Katoh, 2003; Shi *et al.*, 2006).

Breast cancer is a typical example of cancers overexpressing TFF1. As only a low expression is observed in the normal mammary gland (Poulsom *et al.*, 1997;

Deficiency in trefoil factor 1 (TFF1) increases tumorigenicity of human breast cancer cells and mammary tumor development in TFF1-knockout mice

IGV genome viewer



breast cancer~の
Mapping結果

正常ヒト~の
Mapping結果

おまけ

- Moirai
 - 面倒な作業を効率化
 - GUIでの操作
 - クラスターサーバへの対応
 - オリジナルワークフロー作成可能
 - Galaxyと、ちがうの？
 - もうそろそろ、公開されるはず...

RNA-Seq Tools

- http://en.wikipedia.org/wiki/List_of_RNA-Seq_bioinformatics_tools

ありがとうございました