

2015年10月2日  
イルミナサポートウェビナー

# NGSをはじめよう！ BaseSpace で行う RNA-seq 入門 < TopHat/Cufflinks編 >



TopHat Alignment

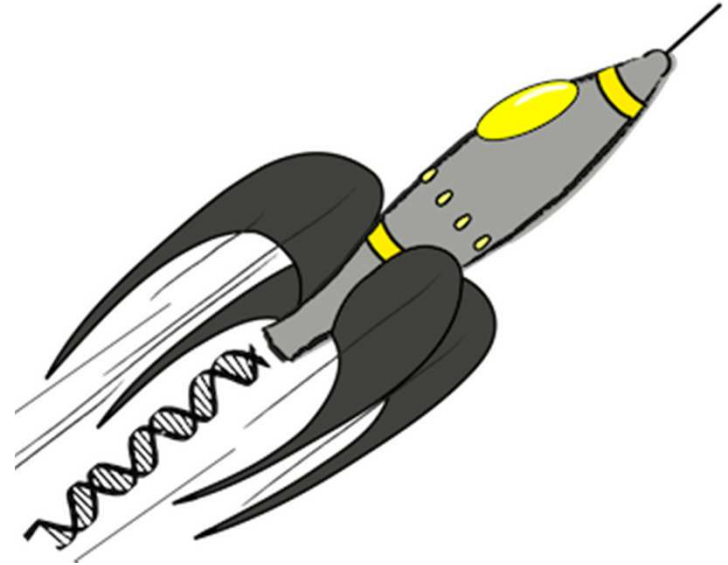


Cufflinks Assembly/  
Diff. Exp

イルミナ株式会社  
バイオインフォマティクス  
サポートサイエンティスト  
癸生川絵里 (Eri Kibukawa)

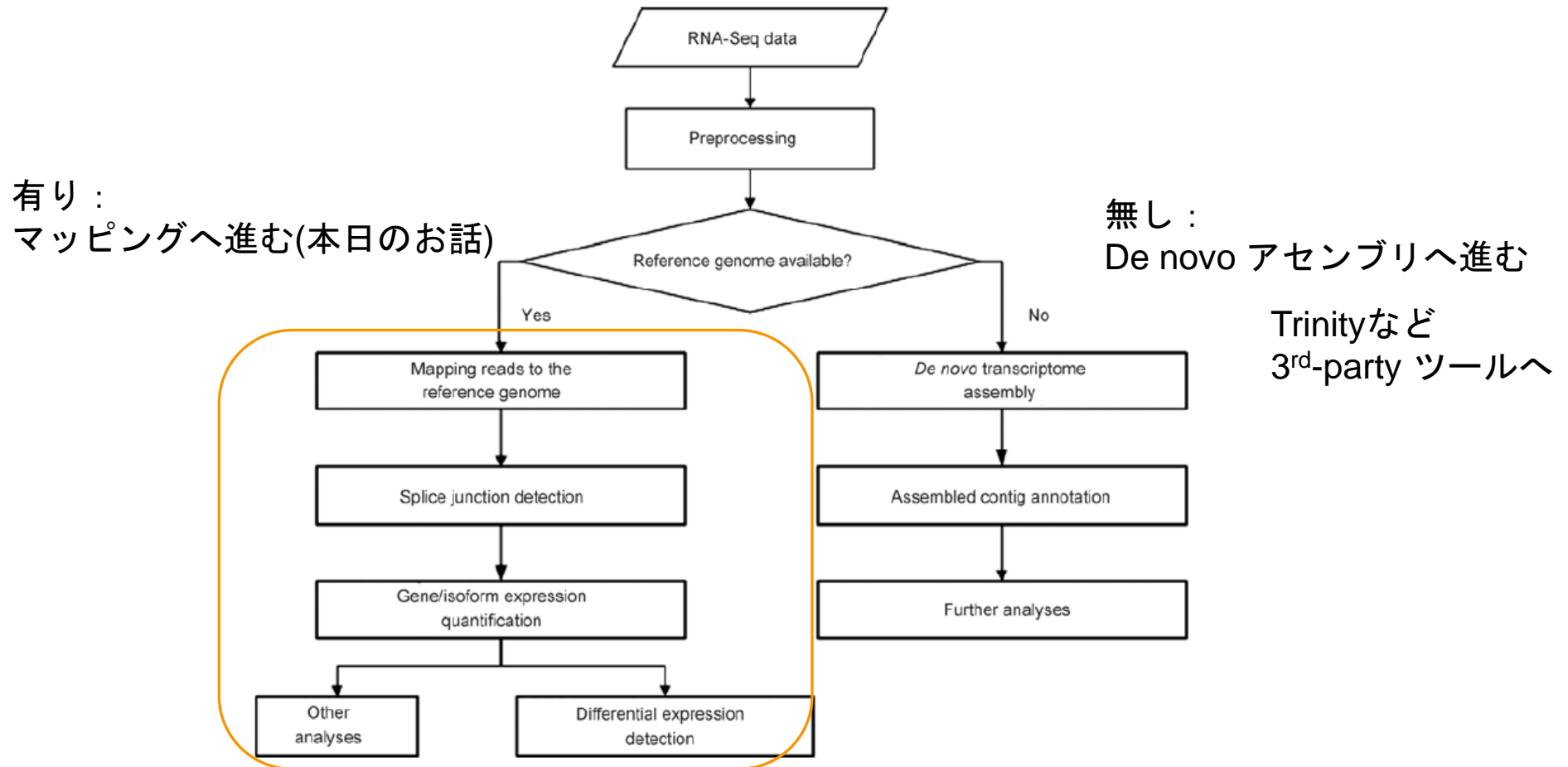
# 本日の内容

- RNA-seq 解析の概要
- BaseSpaceの  
デモデータとRNA-Seq コアアプリ
- TopHatアプリによる解析
- Cufflinks & DEアプリによる解析
- 実験デザインの解析結果への影響



# RNA-seq 解析ワークフロー 概要

->リファレンスゲノムがあるかないか？



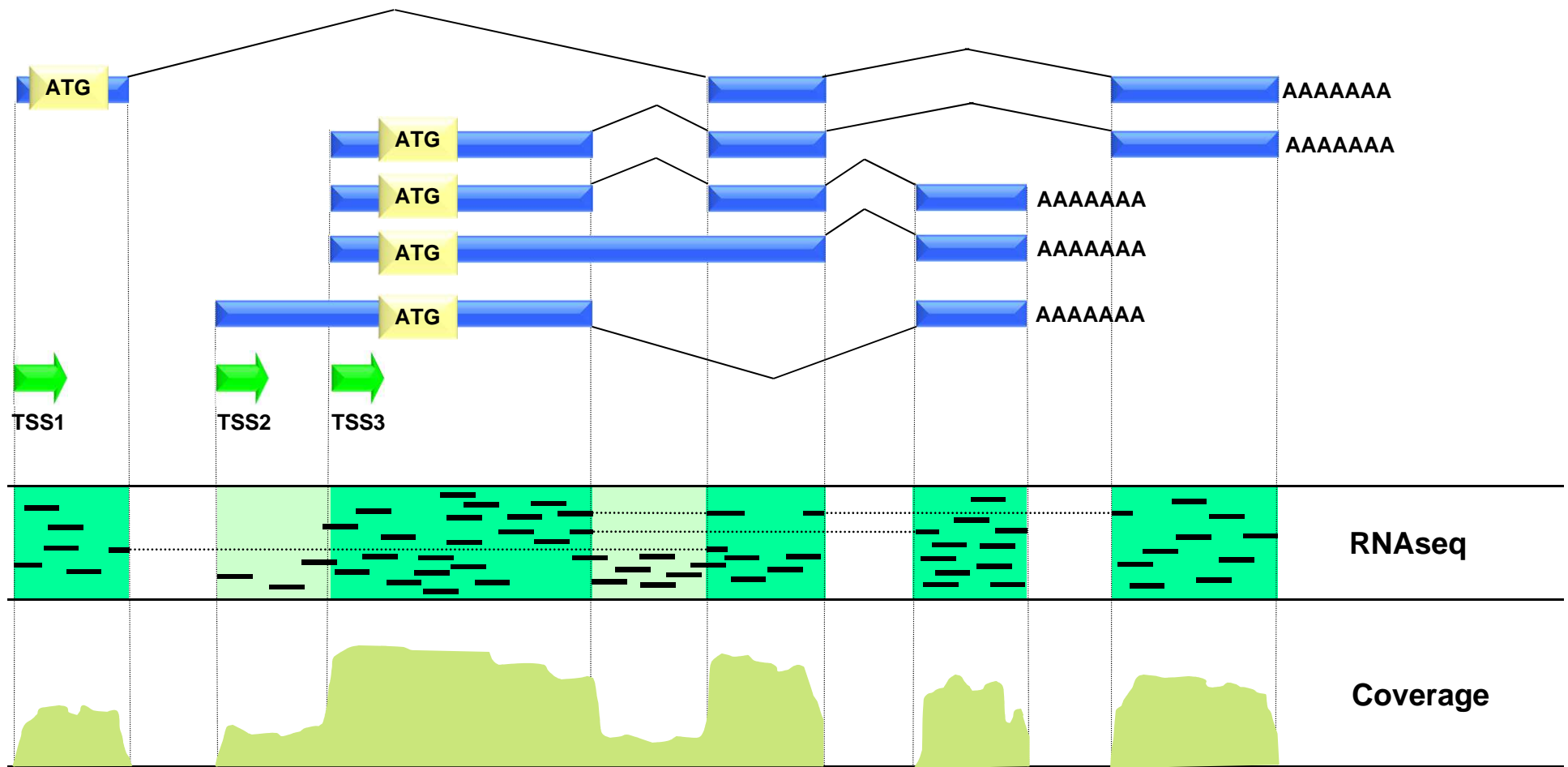
[Sci China Life Sci. 2011 Dec;54\(12\):1121-8. Epub 2012 Jan 7.](#)

# マッピングと発現解析

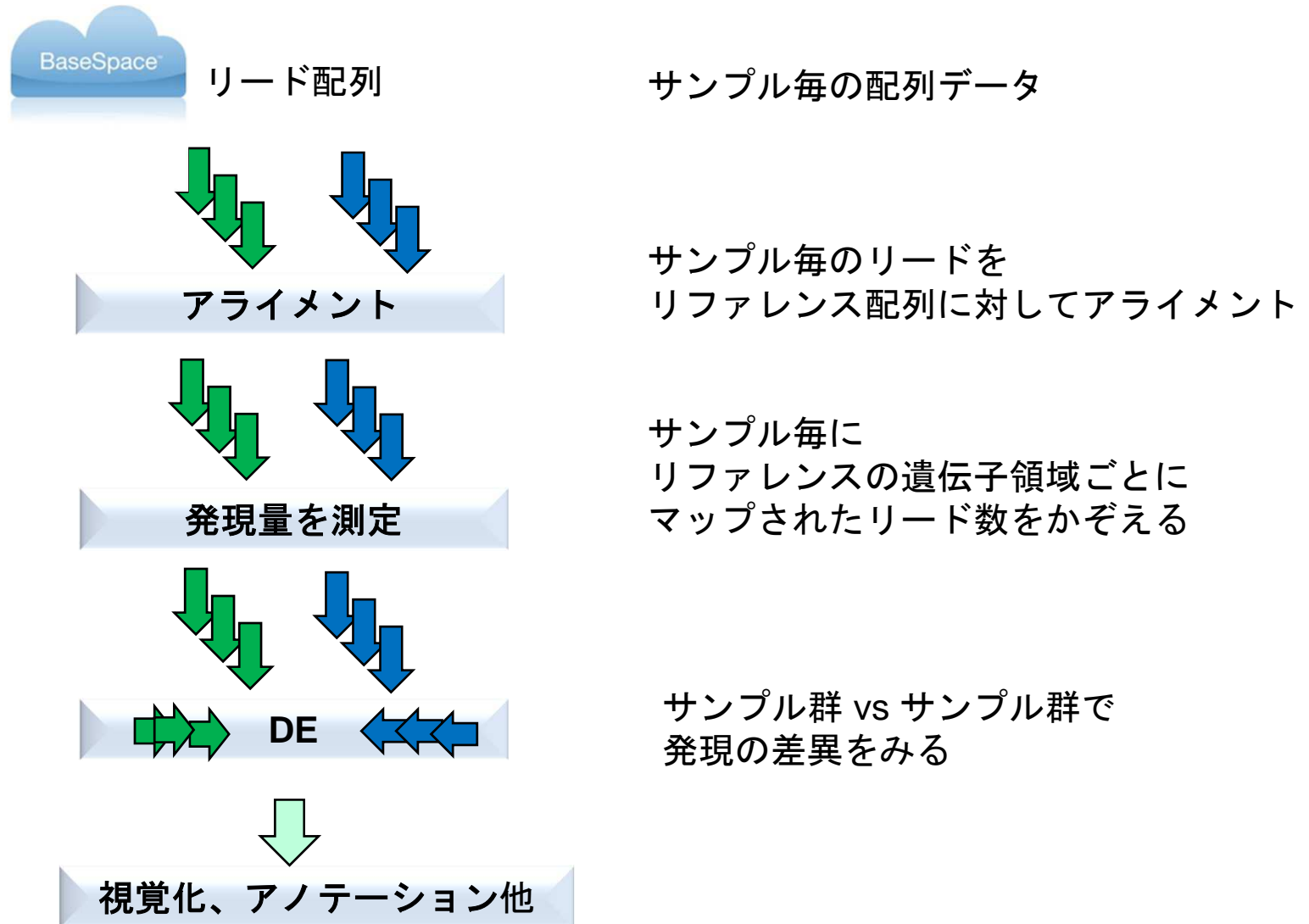
▶特定の遺伝子領域にマップされたリードの数

= 遺伝子転写産物の存在量

に対応していると考える



# RNA-Seq 典型ワークフロー



# RNA Seq アライメント工程 (例)

## アライメント

サンプル毎のリードを  
リファレンス配列に対してアライメント



取り除きたい余剰配列の処理  
rRNA, Mt等



リファレンスゲノムへのアライメント



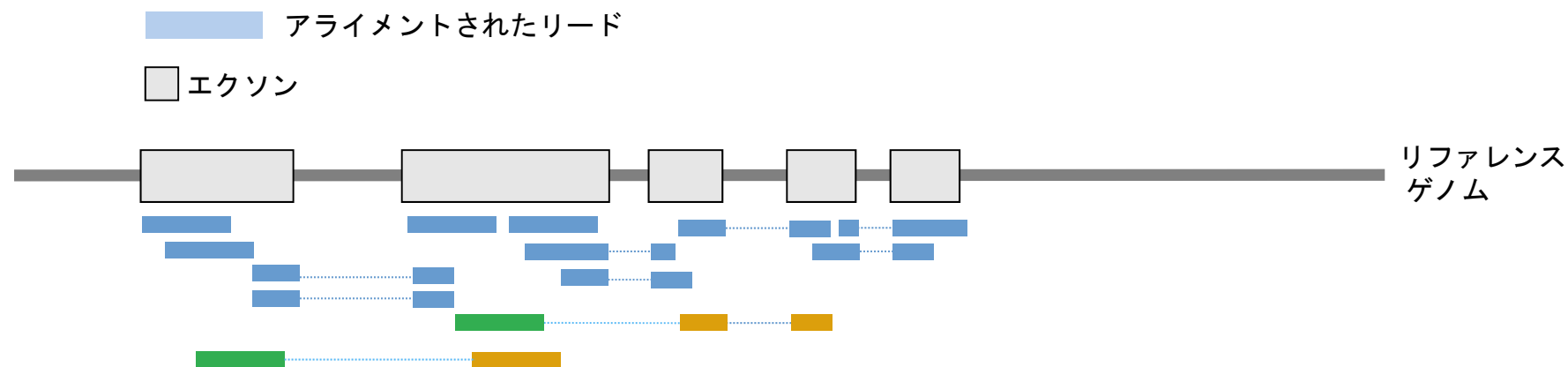
既知スプライスジャンクション  
を考慮したアライメント

# RNA Seq アライメント

## アライメント

サンプル毎のリードを  
リファレンス配列に対してアライメント

アライメントツールは、スプライスジャンクションを考慮したマッピングが必要  
計算量が大きくなるので、ツールによって、それぞれ考慮の仕方の工夫を凝らしている

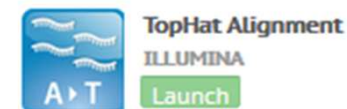


沢山のツールがあり、また計算量が大きい。  
自前でpipelineを構築する場合はしっかりした  
計算機とツール選択やインストールなどの  
メンテナンスが必要。



直ぐに無料で  
お使い頂けます。

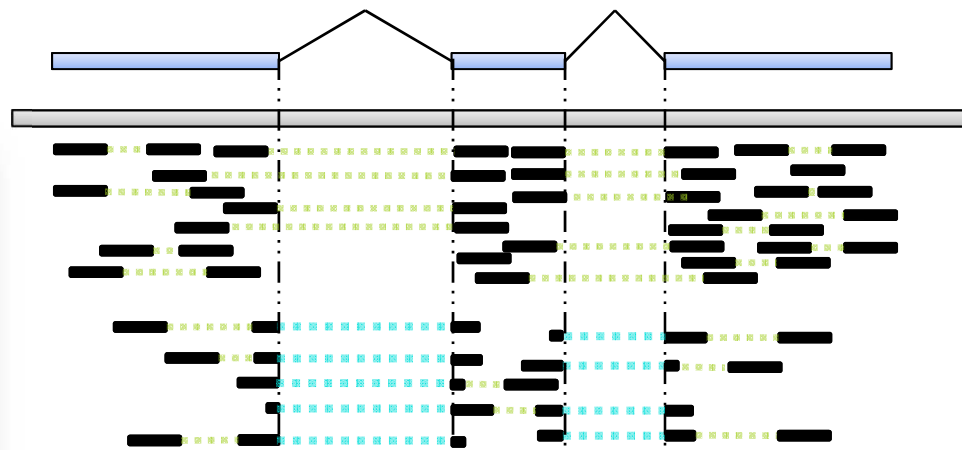
# 例) BaseSpace TopHat アプリの結果 既知遺伝子についての発現リストが得られる



発現リスト => FPKMリスト

Important Files for Download

- Alignments
- Alignment coverage
- Reference FPKM values (genes)**
- Reference FPKM values (transcripts)
- Genome VCF
- TopHat fusion output (20 detected fusions)



**FPKM:** Fragments per kilobase per million mappable (fragments)とは？



# RNA-Seq 発現量の測定と正規化

## RPKM

遺伝子長(全exon長)を1000bpで、リード総塩基数は1Mの場合となるように、  
数えたリード数を標準化する考え

- ▶ RPKM : Reads Per Kilobase of exon per Million of mapped reads

$$\begin{array}{c}
 \text{(対象遺伝子にマップされたリードの塩基数)} \times 1,000 \times 1,000,000 \\
 \hline
 \text{(マップされたリード総塩基数)} \times \text{(対象遺伝子の長さ)}
 \end{array}$$

⇐ 合わせて10の9乗

※ Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer and Barbara Wold Mapping and quantifying mammalian transcriptomes by RNA-Seq Nature Methods, Volume 5, 621 - 628 (2008)

# 遺伝子発現レベルを比較するための正規化の考え方

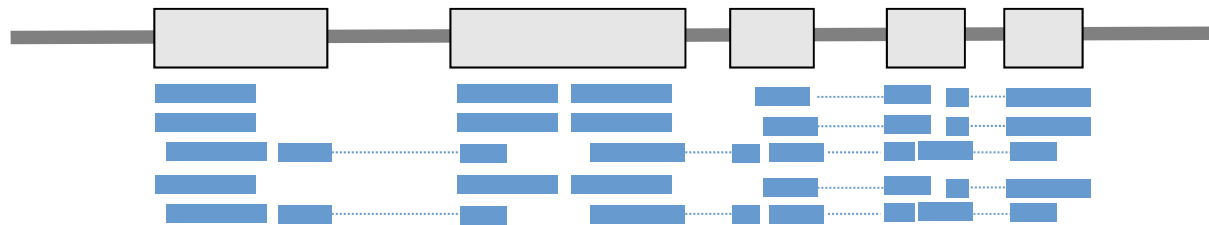
## ① リード総数の影響を考慮する

▶ 発現量の計算はそのサンプルがマップされたリード数、総リード数

(read depth) に影響される

サンプルA

depth = 5 (50 Million 総リード)

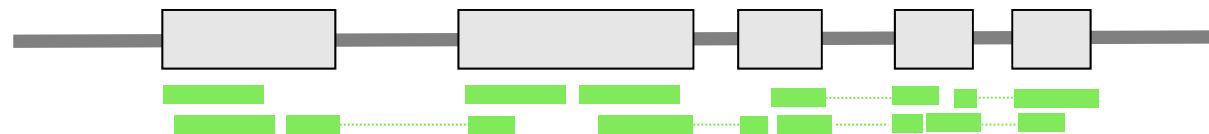


処理サンプル  
 $depth/million\ reads = 5/50 = 0.1$

コントロール  
 $depth/million\ reads = 2/10 = 0.2$

サンプルB (コントロール)

depth = 2 (10 Million 総リード)



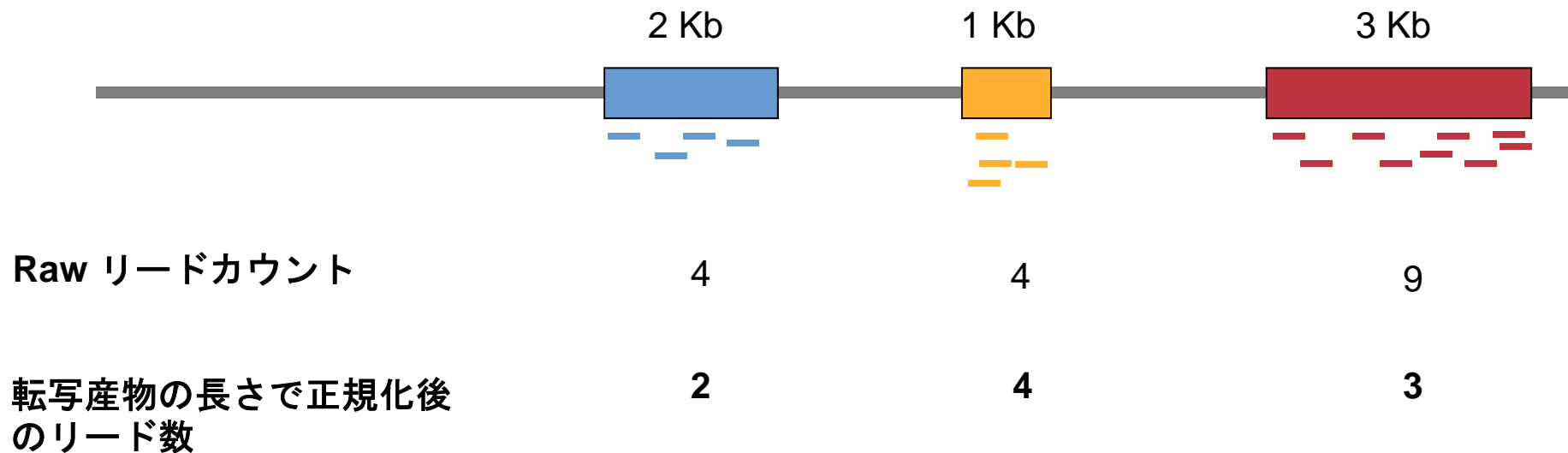
サンプルは  
コントロールに対し  
1/2の遺伝子発現量

# 遺伝子発現レベルを比較するための正規化の考え方

## ② 遺伝子長の影響を考慮する

- ▶ リードカウント数は遺伝子の長さ(全exonの長さ)にも影響される
- ▶ 長ければ長いほどリードがマップされる数が多くなり易い

< 異なる3つの遺伝子を想定 >



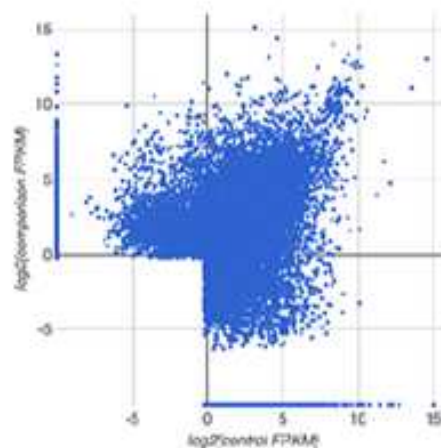
# RNA-Seq 発現差解析



サンプル群 vs サンプル群で  
発現の差異をみる

- ▶ リード数を数え、正規化した数値をもとに  
統計検定を行い発現差異をみていく

サンプル群 B



サンプル群 A

※ 採用統計モデルは使用する  
ツールにより様々であり、  
開発が続けられている



RNA Seqでも複数アプリを搭載しているので  
異なる統計モデルをお試し頂けます

# もっと詳しく知りたい！

## イルミナウェビナー RNA-Seqをはじめよう！シリーズ

[http://www.illumina.co.jp/events/webinar\\_japan.ilmn?ws=ws](http://www.illumina.co.jp/events/webinar_japan.ilmn?ws=ws)

また演者の門田先生のサイトには、より新しく詳細なフォローアップがあり大変参考になります

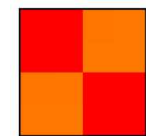
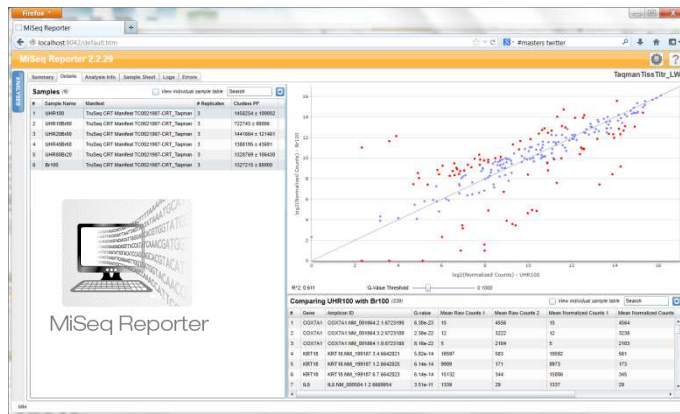
[http://www.iu.a.u-tokyo.ac.jp/~kadota/r\\_seq.html](http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html)



- 門田幸二, 「トランスクリプトームデータ解析戦略2014 (PDF版; YouTube版)」, イルミナウェビナー・RNA-Seqシリーズ, イルミナ株式会社(東京), 2014.07.22  
内容: [イルミナウェビナー](#)にて2011年9月8日と2011年11月17日に行ったRNA-seq周辺のアップデート情報提供がメイン。その後のイルミナウェビナーシリーズを眺めることやアグリバイオインフォマティクスでの私の大学院講義PDF資料のありがたさ(このページのこと)。RNA-seqのおさらい。トランスクリプトーム解析技術(wet側)の進展話。マイクロアレイもヒトトランスクリプトームアレイが出ていること、RNA-seqはIlluminaもshort-readからmedium-readへ。PacBioロングリードのトランスクリプトーム配列を読んだ論文が出始めており、パーソナルゲノムに引き続いてパーソナルトランスクリプトームの時代に来ていることなど。トランスクリプトーム解析技術(dry側)の進展話。遺伝子構造推定系では有名なTophat-Cufflinks/パイプライン以外にも多数のよりよいパイプラインが存在すること。DDBJパイプラインやBaseSpaceを利用することで、Linux-freeでTophat-Cufflinksパイプラインを実行可能であること、しかしそれ以外の多くはLinuxベースであるため、利用したい場合にはLinuxを使いこなせたほうがやはりよいということ。転写物の発現量推定もReXpress、RNA-Skimなどより便利かつ高速に実行できる時代がきていることなど。カウントデータ取得後の発現変動解析はedgeRやDESeqが有名だが、TCCは実質的にiterative edgeRやiterative DESeqに相当するものであり、compcoderRによる客観的な性能評価でも優れていることなど。性能評価に用いたRコードは[20140722\\_compcoderR.txt](#)。1時間分。
- 門田幸二, 「[講義資料](#)」, [アグリバイオインフォマティクス教育研究プログラム](#)の大学院講義科目: [農学生命情報科学特論I](#), 東京大学(東京), 2014.07.02  
内容: [教科書](#)の3.3節と4.3節周辺。マッピングプログラムは大きくbowtieなどのbasic aligner (unspliced aligner)とtophatなどのsplice-aware aligner (spliced aligner)に大別されること。splice-aware alignerの基本的なイメージ。ゲノム配列既知の場合の遺伝子構造推定としてTophat-Cufflinks/パイプラインの基本形を紹介。既知遺伝子(または転写物)の発現解析でよい場合は、トランスクリプトーム配列へのマッピングでよい。最近ではSailfishやRNA-Skimなど、k-merに基づくalignment-freeな方法が目目されていることなど。研究目的別留意点として、遺伝子間比較の場合とサンプル間比較の場合、配列長補正、総リード数補正、RPKMなど。長い転写物ほどマップされるリード数が多い傾向をRで確認。GSE42212のヒトRNA-seqデータのFASTQファイル取得以降の一通りの解析。実際に行ったのは、カウントデータ取得以降のTCCパッケージを用いたサンプル間クラスタリング、発現変動遺伝子(DEG)同定。M-A plotのおさらい。結果の解釈。FDR、分布やモデルの説明。倍率変化でDEG同定を行う場合との比較。2コマ(2×90 min)分。

# 可視化やアノテーション

# 可視化、アノテーション他



Filters



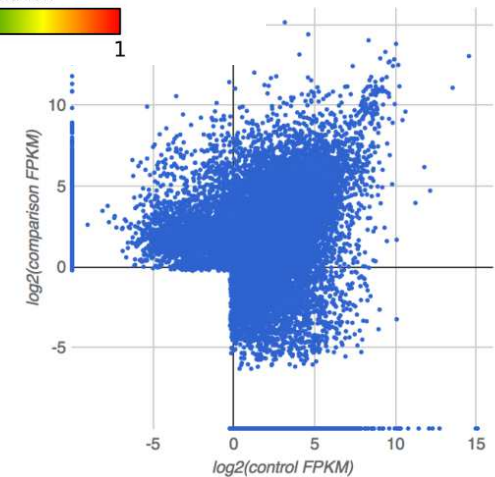
Significant

Choose a value...

Status

OK

Gene



The screenshot shows the BaseSpace ONSITE analysis report for the gene NEUROD1. It includes sections for 'Most Correlated Tissues', 'Most Correlated Diseases', 'Most Correlated Compounds', and 'Most Correlated Gene Perturbations'.

- Most Correlated Tissues:**
  - Cerebellar hemisphere
  - Cerebellar vermis
  - Cerebellum
  - Cerebellum peduncles
  - Pons
- Most Correlated Diseases:**
  - Diabetes mellitus type 1
  - Diabetes mellitus
  - Disorder of endocrine pancreas
  - Allergic disorder
  - Measles
- Most Correlated Compounds:**
  - Chronic Acid
  - Thioctic Acid
  - Tetanus Toxin
  - Sumatriptan
  - Cholera Toxin
- Most Correlated Gene Perturbations:**
  - NEUROD1
  - SETDB1
  - FBX1
  - SLC11A1
  - HR

Test ID	Gene	Locus	Status	log <sub>2</sub> (control FPKM)	log <sub>2</sub> (comparison FPKM)	log <sub>2</sub> (Ratio)	q Value	Significant
XLOC_000987	-	chr1:152020810-152021644	OK	-10	2.99	-12.99	0.164383	x
XLOC_001014	S100A9	chr1:153330329-153333503	OK	4.61	14.4	-9.79	0.582102	x
XLOC_001015	-	chr1:153359119-153359585	OK	-10	2.23	-12.23	0.22233	x
XLOC_001017	S100A1	chr1:153591275-153618799	OK	7.09	-10	17.09	0.264328	x
XLOC_001018	CHTOP	chr1:153591275-153618799	OK	4.99	4.67	0.32	0.954448	x
XLOC_001019	SNAPIN	chr1:153631120-153643504	OK	4.61	4.11	0.5	0.957225	x

<http://res.illumina.com/documents/products/technotes/technote-basespace-rna-seq.pdf>

# その他ツールのリストアップ

<http://seqanswers.com/wiki/Software/list>

<http://seqanswers.com/wiki/RNA-Seq>

The bioinformatics applications assigned the Biological domain **RNA-Seq** (topic\_3170 @) are tabulated below.

**Definition:**

- A topic concerning high-throughput sequencing of cDNA to measure the RNA content (transcriptome) of a sample, for example, to investigate how different alleles of a gene are expressed, detect post-transcriptional mutations or identify gene fusions.

**Synonyms:**

- WTSS
- Small RNA-Seq
- Whole transcriptome shotgun sequencing
- RNA-seq
- Small RNA-seq

Query returned 46 results.

	Biological domain	Bioinformatics method	Input format	Output format
ArrayExpressHTS	<b>RNA-Seq</b> RNA-Seq Quantitation		FASTQ	
Avadis NGS	ChIP-Seq DNA-Seq <b>RNA-Seq</b> Small RNA Pathway analysis	Alignment Quality Control Sequence analysis Visualization Biological Contextualization	SAM BAM BED ELAND FASTA FASTQ	
Chipster	ChIP-Seq <b>RNA-Seq</b> MiRNA-Seq MeDIP-Seq	QC Filtering Trimming Mapping Peak calling Motif detection Differential expression Pathway analysis Methylation analysis Genomic region matching Genome browser	FASTQ SAM BAM BED GTF	FASTQ SAM BAM BED GTF
	Genomics Whole Genome Resequencing	Mapping Assembly Alignment Colorspace	FASTA	FASTA FASTQ GFF

# 様々なサードパーティーツールのご紹介

可視化、アノテーション他

[http://res.illumina.com/documents/products/datasheets/datasheet\\_rnaseq\\_analysis.pdf](http://res.illumina.com/documents/products/datasheets/datasheet_rnaseq_analysis.pdf)



Table 1: RNA-Seq Analysis Tools

Tool or Suite	Description	Availability	Link
Galaxy	Free form-based access to Bowtie, TopHat, and Cufflinks. Web browser client.	Academic/Open Source	galaxy.psu.edu
GenePattern	Free form-based access to Bowtie, TopHat, and Cufflinks. Local client.	Academic/Closed Source	www.broadinstitute.org/cancer/software/genepattern
Partek	Advanced statistics and interactive visualization for microarray and sequencing data.	Commercial	www.partek.com
CLC Bio	Software analyzing and visualizing sequencing data	Commercial	www.clcbio.com
GeoSpiza	Cloud-based analysis for microarray and sequencing data	Commercial	www.geospiza.com
GenomeQuest	Software for sequence data management	Commercial	www.genomequest.com
Avadis NGS	Software for sequence data analysis and management	Commercial	www.avadis-ngs.com
Ingenuity IPA	Software for biological pathway analysis	Commercial	www.ingenuity.com/products/pathways_analysis.html





# イルミナ RNA-Seqワークフローの例

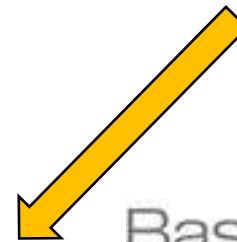
TruSeq® RNA



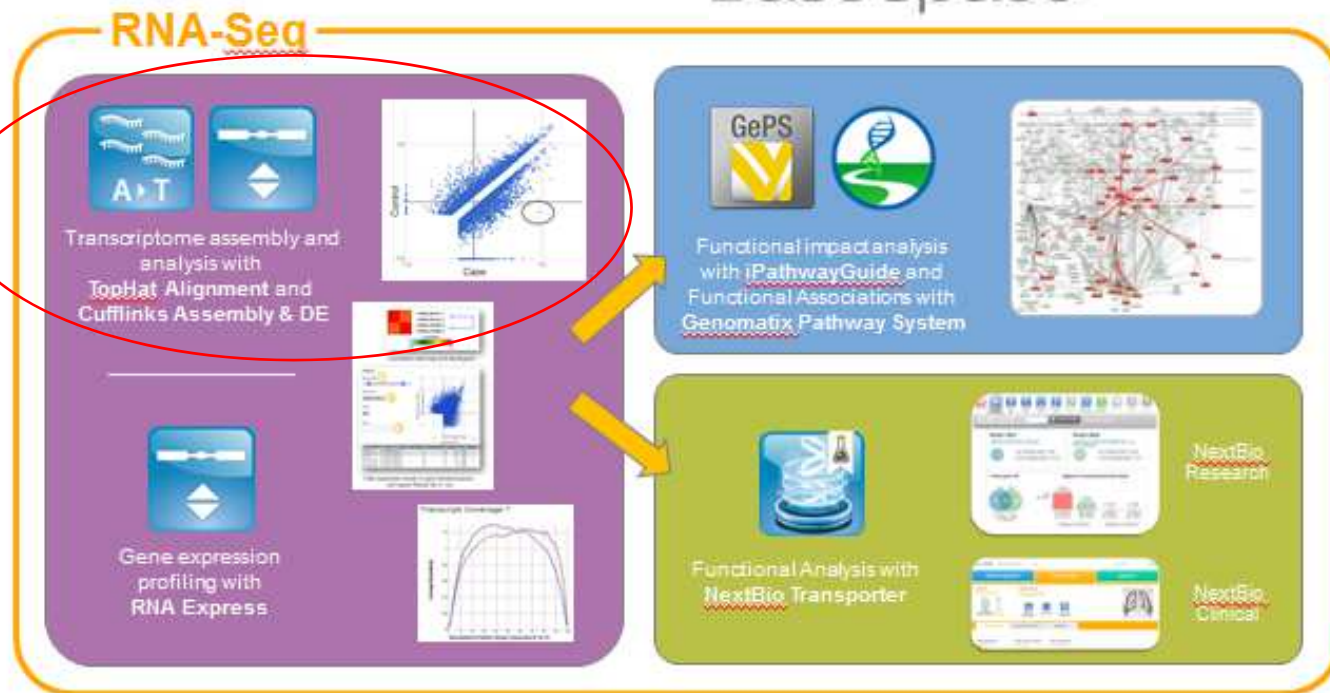
**NeoPrep**  
サンプル自動調整



イルミナシーケンサー  
リアルタイムモニタリング



BaseSpace®



他社製  
アプリ

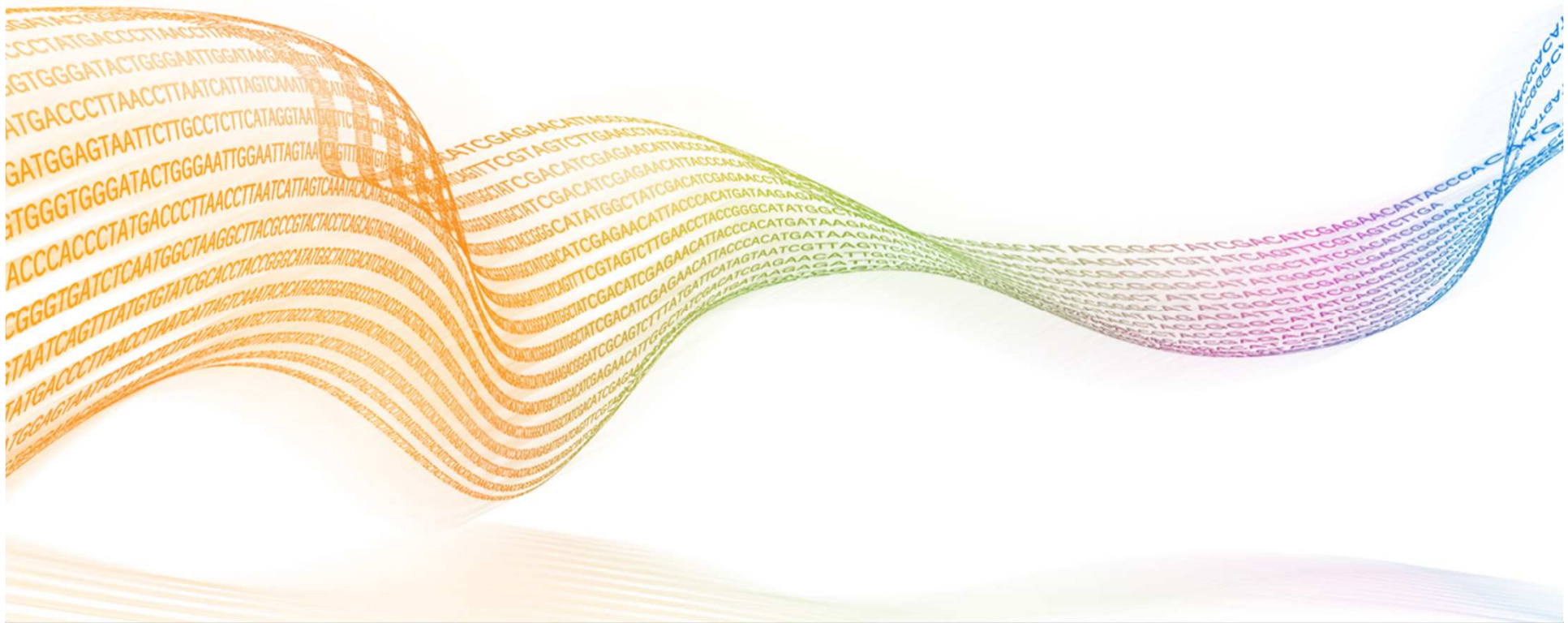
イルミナ  
ラボアプリ

イルミナ  
コアアプリ

BaseSpace®



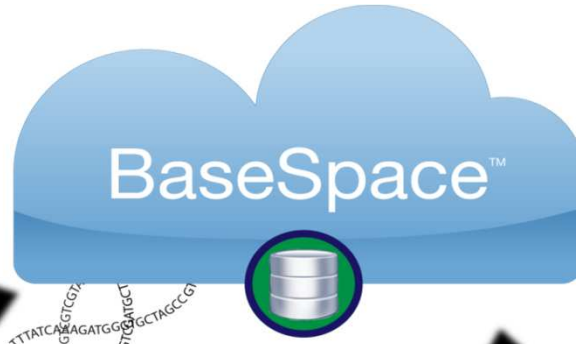
# BaseSpaceの Public Data(公開デモデータ)とアプリ



# BaseSpace 概要

指定のワークフロー自動実行  
ユーザによるアプリの選択実行

逐次アップロードにより、  
FASTQ自動生成 開始が  
ゼロダウンタイム



データ保全  
データ管理  
削除やフォルダ  
の追加

共同研究者とのシェア  
権限移譲



装置からシームレスに接続、  
BSでリアルタイムモニタリング可能

アプリをブラウザから楽に実行。  
解析パラメータやアプリを変更し簡単に  
解析を再実行。メール通知を待つだけ。

<http://basespace.com> からログイン  
初めてのの方はSign up から簡単に登録いただけます。

BaseSpace®  
Genomics Cloud Computing

illuminatm

Sign up Log in

Now Available on the BaseSpace AppStore

RNA-Seq Workflow

Filters

|log2(ratio)|  
0.0 45.0

Significant  
Choose a value...

Status  
OK

Gene

Coverage (Normalized)

Normalized Position Along Transcript (5' to 3')

log2(comparison FPKM)

log2(control FPKM)

TopHat Alignment

Cufflinks Assembly & Differential Expression

ログイン

# 公開デモデータ (BaseSpace CloudのPublic Dataにあり)

The screenshot shows the BaseSpace Cloud interface. The top navigation bar includes 'BaseSpace', 'Dashboard', 'Prep', 'Runs', 'Projects', 'Apps', 'Public Data' (circled in red), and 'Help'. Below the navigation bar, there are three data entries:

- > HiSeq 2500: TruSeq Stranded mRNA LT (SEQC: UHR & Brain)  
RNA-Seq
- > HiSeq 2000: TruSeq Stranded Total RNA (MAQC)  
RNA-Seq | Differential Expression
- ▼ NextSeq 500: RNA-Seq (8plex)  
NextSeq 500 data generated for reference human brain RNA and universal human reference RNA (UHRR). Libraries were prepared using TruSeq stranded mRNA or TruSeq Total RNA with Ribo-Zero Gold reagent kits.  
RNA-Seq

Below the 'NextSeq 500: RNA-Seq (8plex)' entry, there is a table with two rows and two columns:

Run	NextSeq 500: RNA-Seq (8plex) (49.41 GB)	Import
Project	NextSeq 500: RNA-Seq (8plex) 0	Import

The 'Import' button for the Project row is circled in red. A white arrow points from this button to the text below.

On the right side of the interface, there are two sections:

- Research Areas**
  - Cancer Research
  - Genetic Disease
  - Complex Disease
  - Microbial Research
- Categories**
  - Exome
  - Resequencing
  - Small RNA
  - Targeted Sequencing
  - De Novo Assembly
  - RNA-Seq (circled in red)
  - Gene Fusion Detection
  - ChIP-Seq
  - Methyl-Seq
  - Metagenomics
  - Tumor Normal
  - Variant Analysis
  - Differential Expression
  - Quality

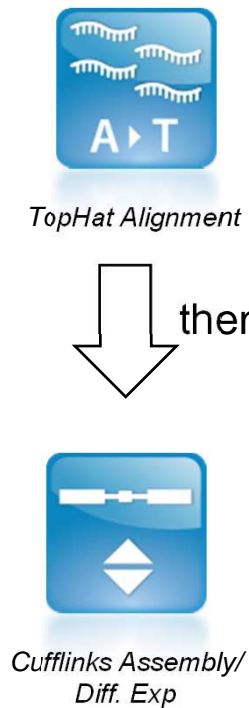
Projectをインポートして頂きますと、既に終了している解析結果から、レポート等を見ていただけるとともに、実際にアプリを実行いただくことも可能です。

# イルミナはBaseSpace 上にRNA Seq 用の 3つのアプリをご提供しています

手法	BaseSpace アプリ	アイコン	内容
RNA-Seq	TopHat アライメント		<ul style="list-style-type: none"> <li>業界標準のTopHat2を使ったRNA-Seqアライメントとカウンティング</li> <li>融合遺伝子のコール(オプション)</li> <li>ISAAC Variant Callerを使ったcSNPコール</li> <li>結果はCufflinks Assembly &amp; DE Appでさらに解析可能</li> </ul>
	Cufflinks アセンブル & 遺伝子発現解析		<ul style="list-style-type: none"> <li>詳細遺伝子発現差解析</li> <li>選択的転写産物のアセンブルと新規転写産物予測</li> </ul>
	RNAExpress		<ul style="list-style-type: none"> <li>迅速な遺伝子発現プロファイルをSTARアライメントとDESeq2で実現</li> <li>遺伝子レベルの遺伝子発現に特化</li> </ul>

- リファレンスは現在 hg19, mm10, rn5
- 遺伝子構造は、RefSeqとGENCODEを選択可
- カスタムリファレンスはアップロードできない (2015/10現在)

# アプリの違い：2通りの使い方

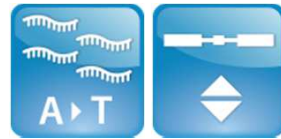


QC工程をはさめる！



発現解析(Expression)をノンストップ特急 (Express)で！

# アプリの違い：主要機能面



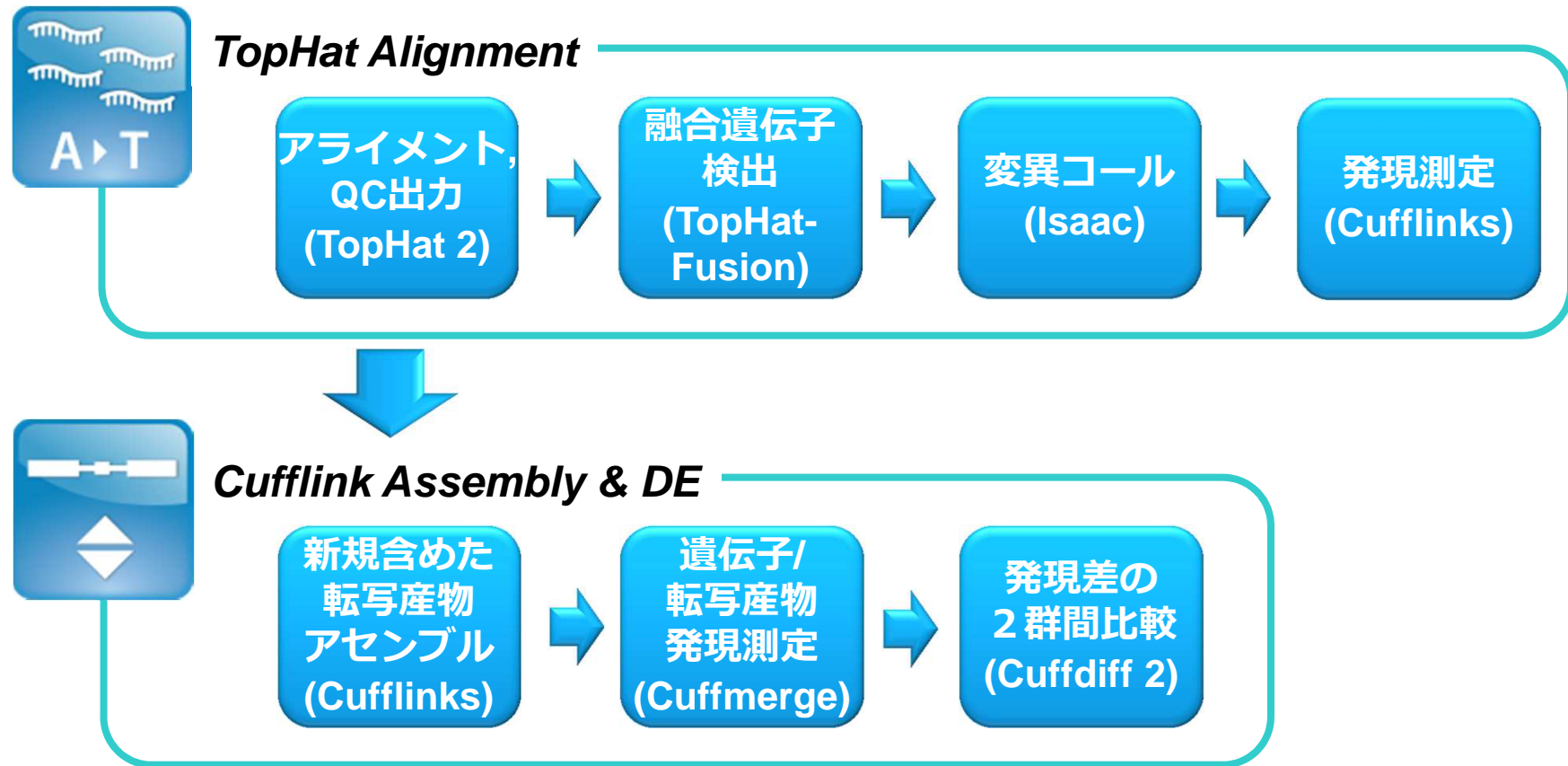
機能	TopHat/ Cufflinks	RNA Express
○ シーケンス量のフィルター	あり	あり
○ シーケンスアライメント	あり	あり
変異コール	あり	なし
融合遺伝子コール	あり	なし
転写産物アセンブル	あり	なし
遺伝子量予測	あり	なし
転写産物量予測	あり	なし
○ 遺伝子発現差の解析	あり	あり

高機能、詳細解析の実行

かんたん、速い



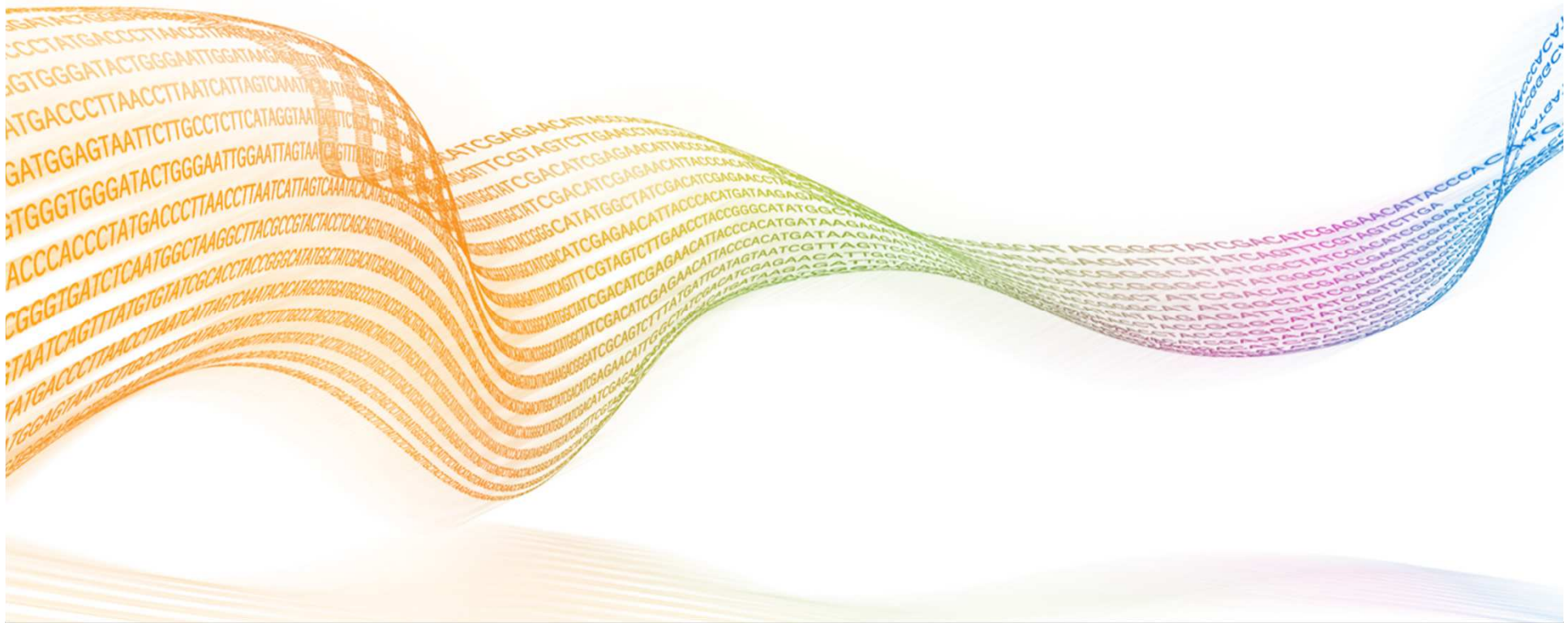
# TopHat Alignment + CufflinksAssembly&DE による ワークフローの内包ツール



使用ソフトウェアのバージョン

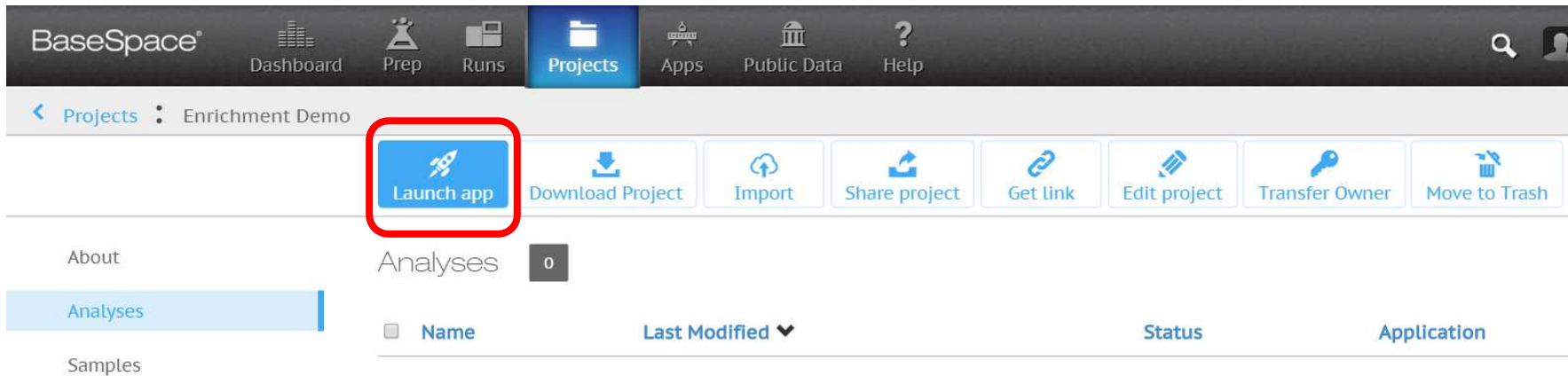
TopHat2 v2.0.7, Bowtie 0.12.9, Cufflinks 2.1.1, Isaac Variant Caller 2.0.5, Picard tools 1.72

# BaseSpace TopHat Alignment アプリ

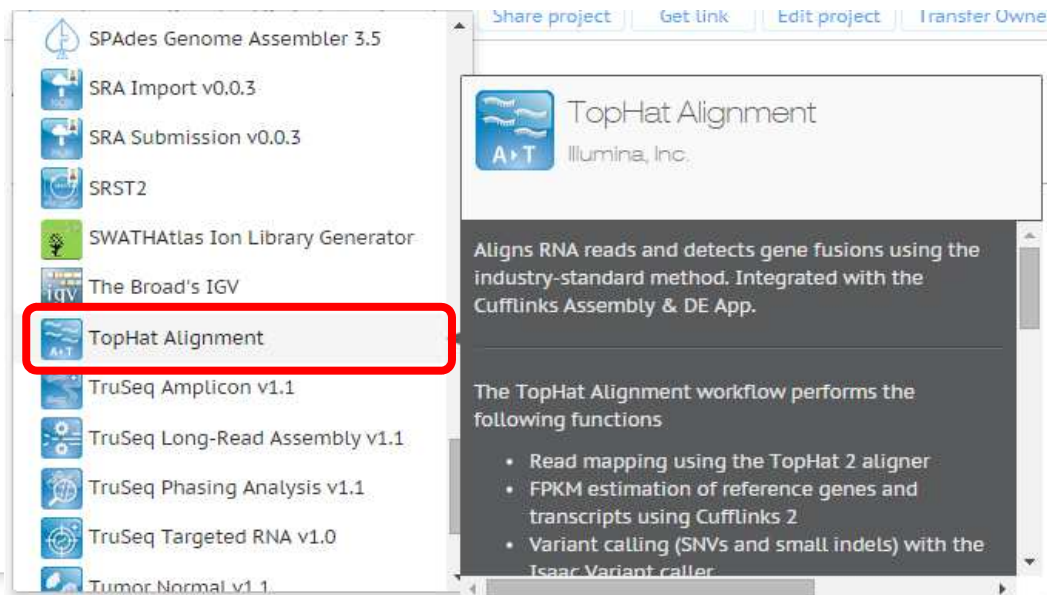


# TopHat Alignmentの実行

① Project WindowでLaunch Appを選択する



② TopHat Alignmentの選択



# TopHat Alignmentの実行 設定画面



TopHat Alignment

App Session Name: TopHat Alignment 05/16/2015 8:13:50

Save Results To: Select Project(s):  
Transcriptome Demo

Samples: Select Sample(s):

Select All	Stranded	
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
mRNA-Brain-C4	<input checked="" type="checkbox"/>	x
mRNA-UHRR-C2	<input checked="" type="checkbox"/>	x
mRNA-UHRR-C1	<input checked="" type="checkbox"/>	x
mRNA-Brain-C6	<input checked="" type="checkbox"/>	x

サンプルの選択を行う

Strandedでサンプル調整された場合は忘れずチェック

Reference Genome: Homo sapiens/hg19 (RefSeq)

リファレンスゲノムと遺伝子モデルを選択

## Options

Call Fusions:

Trim TruSeq Adapters:

オプションの設定

- Homo sapiens/hg19 (RefSeq)
- Homo sapiens/hg19 (RefSeq)**
- Homo sapiens/hg19 (Gencode)
- Mus musculus/mm10 (RefSeq)
- Rattus norvegicus/rn5 (RefSeq)

# TopHatアプリの実行結果： genes/transcriptsのFPKMリスト



TopHat Alignment

## FPKMリスト

	A	D	E	F	G	H	I	J	L	M
	tracking	gene_id	gene_short_name	tss_id	locus	length	coverage	FPKM	conf_hi	FPKM_conf_us
1	tracking	OR4F5	OR4F5	TSS14428	chr1:69090-7-	-	-	0	0	0 OK
2	OR4F5	FAM138A	FAM138A	TSS8403	chr1:34610-3-	-	-	0.193449	0.3624	OK
3	FAM138A	DDX11L1	DDX11L1	TSS14844	chr1:11873-1-	-	-	0.140593	0.0808	OK
4	DDX11L1	WASH7P	WASH7P	TSS7514	chr1:14361-2-	-	-	2.8629	2.4459	OK
5	WASH7P	LOC729737	LOC729737	TSS18541	chr1:134772-	-	-	0.118328	0.4744	OK
6	LOC729737	OR4F29	OR4F29	TSS12680	chr1:621095-	-	-	0.0893469	0.1996	OK
7	OR4F29	OR4F29	OR4F29	TSS4943	chr1:367658-	-	-	0.232799	0.1996	OK
8	OR4F29	LOC1001322	LOC1001322	TSS12303	chr1:323891-	-	-	2.55131	0.91286	OK
9	LOC1001322	LOC1001333	LOC1001333	TSS12303	chr1:323891-	-	-	1.69886	0.79408	OK
10	LOC1001333	LOC1001333	LOC1001333	TSS13053	chr1:661138-	-	-	1.36847	0.08071	OK
11	LOC1001333	LOC1002880	LOC1002880	TSS8709	chr1:700244-	-	-	10.8821	0.2079	OK
12	LOC1002880	LINC00115	LINC00115	TSS18312	chr1:761585-	-	-	0.153922	0.1458	OK
13	LINC00115	LOC1001304	LOC1001304	TSS20197	chr1:852952-	-	-	0.0826353	0	OK
14	LOC1001304	FAM41C	FAM41C	TSS20841	chr1:803450-	-	-	0.151429	0.6164	OK
15	FAM41C	KLHL17	KLHL17	TSS17580	chr1:895966-	-	-	0.272212	0.3658	OK
16	KLHL17	PLEKHN1	PLEKHN1	TSS12072	chr1:901876-	-	-	0	0	OK
17	PLEKHN1	C1orf170	C1orf170	TSS8609	chr1:910578-	-	-	1.01169	0	OK
18	C1orf170	ISG15	ISG15	TSS16361	chr1:948846-	-	-	3.46688	0.21552	OK
19	ISG15	HES4	HES4	TSS26547	chr1:934341-	-	-	1.22598	0.57888	OK
20	HES4	RNF223	RNF223	TSS16029	chr1:100712-	-	-	0	0	OK
21	RNF223	LOC1002880	LOC1002880					10.9593		

TopHatアプリの発現測定ポイント；

- ☆ 既知のご指定の(RefSeq or GENCODE) genes, transcriptsモデルのみでのカウント集計。
- ☆ 新規に予測しそれらのカウントなどは、Cufflinks&DEアプリが対応している。

# TopHat Alignmentの実行結果



TopHat Alignment

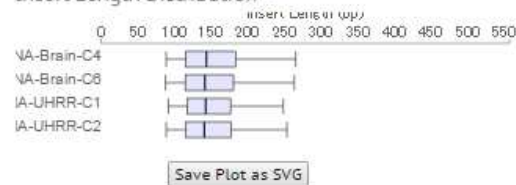
App Session TopHat Alignment 05/16/2015 8:13:50

Summary [i](#)

	Reads	Number of Reads	% Total Aligned	% Abundant	% Unaligned	Median CV Coverage Uniformity	% Stranded
mRNA-Brain-C4	75/75	97,730,535	92.04%	15.69%	7.96%	0.55	99.36%
mRNA-Brain-C6	75/75	94,064,211	95.99%	15.21%	4.01%	0.55	99.11%
mRNA-UHRR-C1	75/75	83,374,339	96.31%	11.21%	3.69%	0.55	99.47%
mRNA-UHRR-C2	75/75	84,897,013	96.69%	9.99%	3.31%	0.54	99.46%

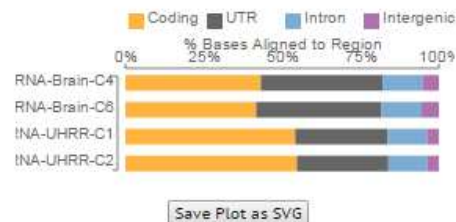
**Summary**  
アライメント割合、カバレッジの均一性など

Insert Length Distribution [i](#)



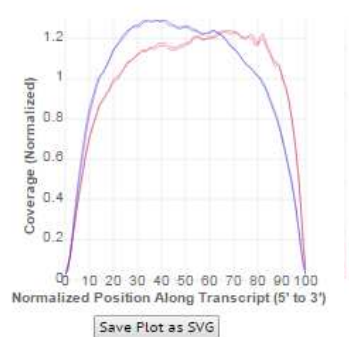
**Insert Length Distribution**  
ライブラリーインサートのサイズ分布

Alignment Distribution [i](#)



**Alignment Distribution**  
Coding Exon, UTR, Intron, Intergenicへのアライメント割合

Transcript Coverage [i](#)



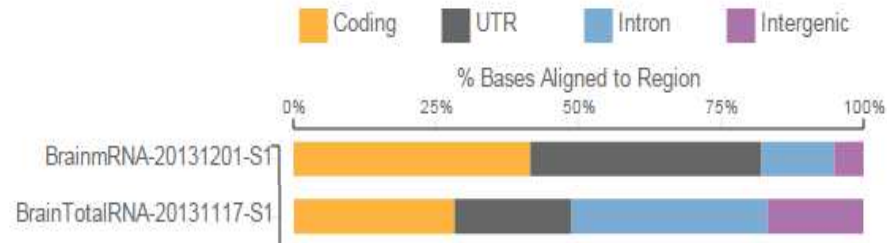
**Transcript Coverage**  
転写産物の5'末端から3'末端までのカバレッジの分布

# TopHat Alignmentの実行結果



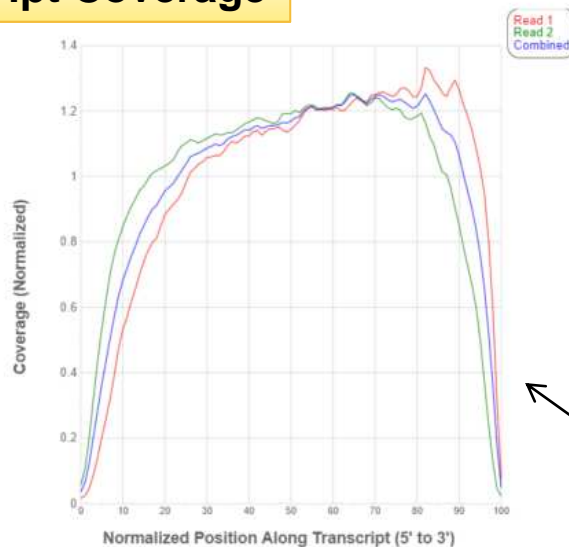
TopHat Alignment

## Alignment Distribution



Total RNAキットを用いた場合には  
IntronとIntergenic RegionへAlignment  
されるリードの割合が増える

## Transcript Coverage



mRNAキットを用いた場合には、PolyA  
RNAの単離を行っている。RNAが分解し  
ている場合には5'側のカバレッジが低く  
なる

いわゆる3' bias

もし問題があればライブラリー調製やRNA抽出の問題が考えられる。  
-> 続くCufflinks&DEアプリの解析からは外し、実験へフィードバックするなど。

# TopHat Alignmentの実行結果



TopHat Alignment

## Variant Calls <sup>i</sup>

Homozygous reference	34,651,888
Heterozygous	62,815
Homozygous variant	10,439
SNV	69,593
Indel	3,710
T <sub>n</sub> /T <sub>v</sub>	3.30

変異コールの結果

## Important Files for Download

<a href="#">Alignments</a>
<a href="#">Alignment coverage</a>
<a href="#">Reference FPKM values (genes)</a>
<a href="#">Reference FPKM values (transcripts)</a>
<a href="#">Genome VCF</a>
<a href="#">TopHat fusion output (13 detected fusions)</a>

この段階での発現解析結果(FPKM)と変異解析結果(gVCF)もダウンロード可能

融合遺伝子の検出結果



# TopHatアプリの出力結果： 変異解析の結果



TopHat Alignment

## Important Files for Download

Alignments

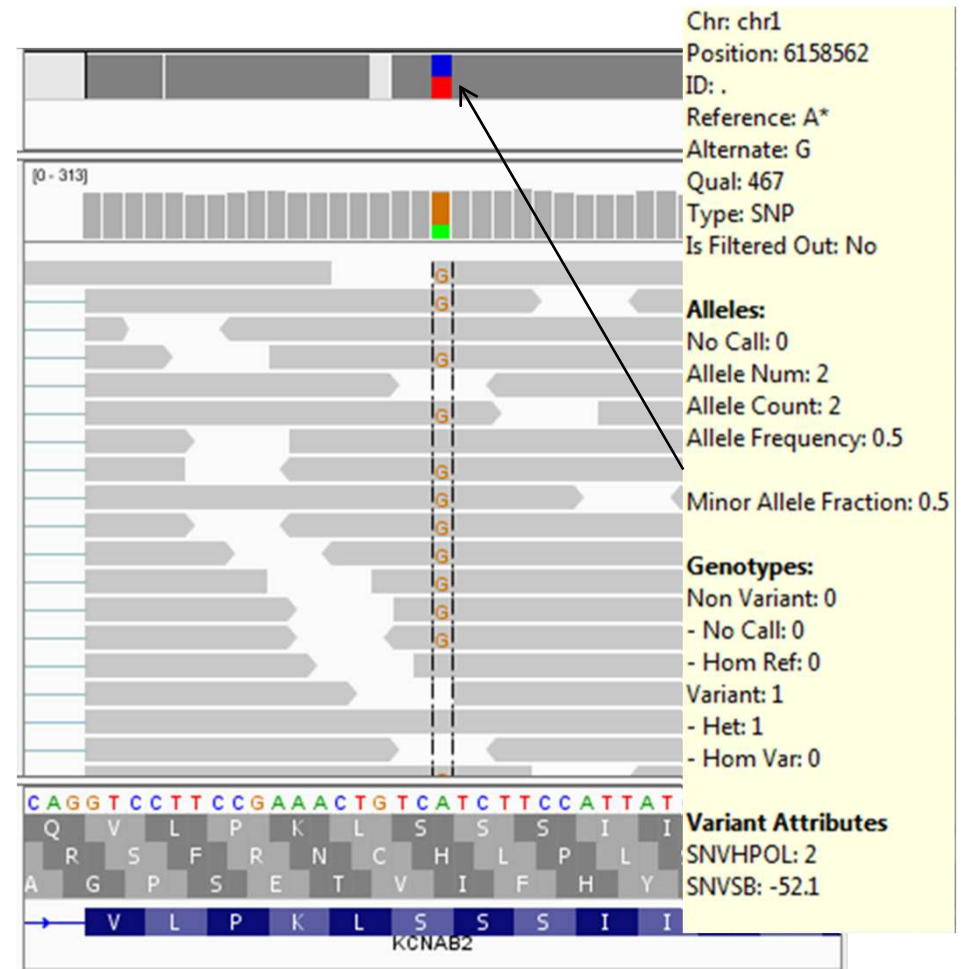
Alignment coverage

Reference FPKM values (genes)

Reference FPKM values (transcripts)

Genome VCF

TopHat fusion output (20 detected fusions)



.gvcf

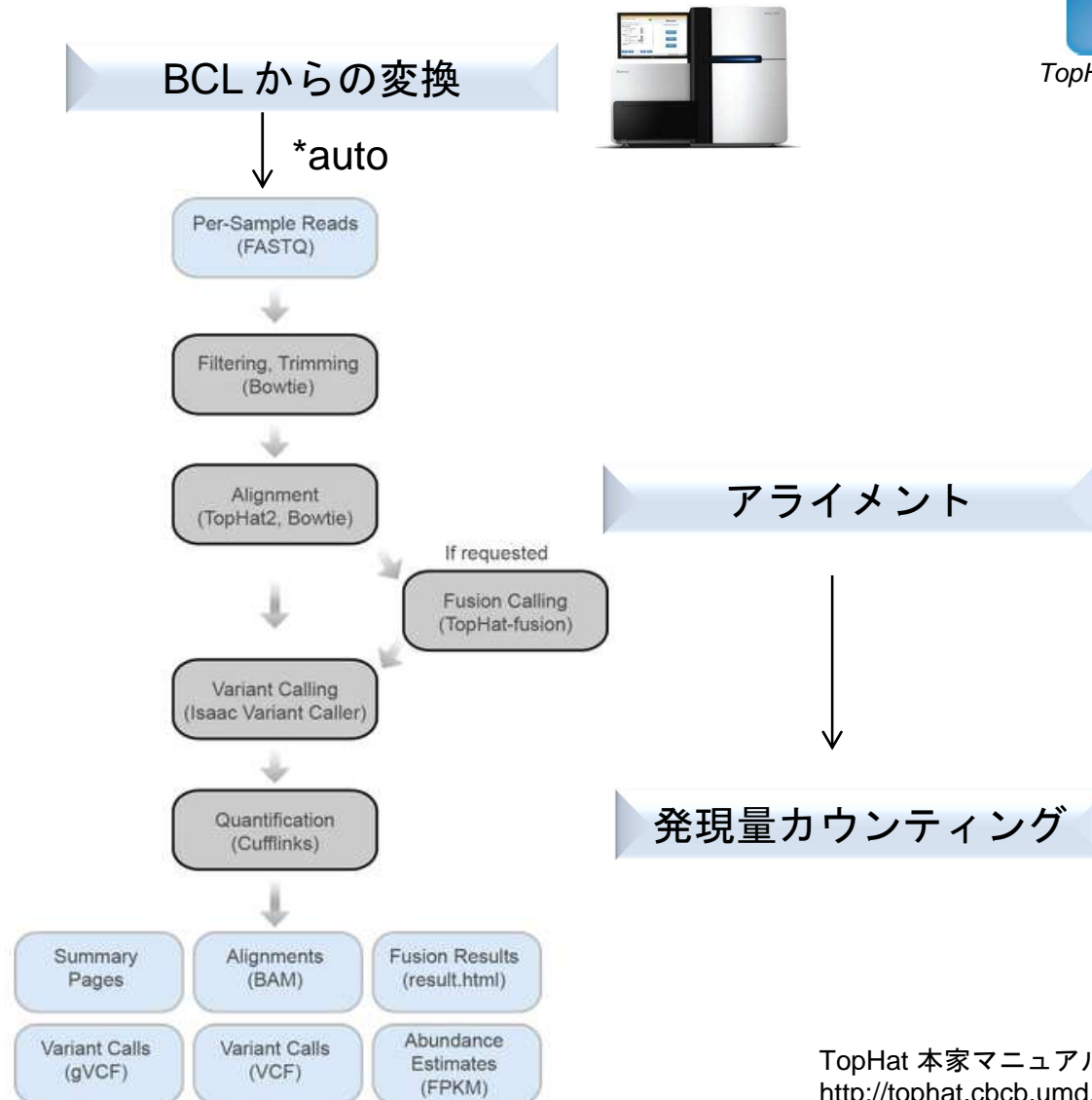
(<https://sites.google.com/site/gvcftools/home/about-gvcf>)



# TopHat Alignment App のパイプライン



TopHat Alignment

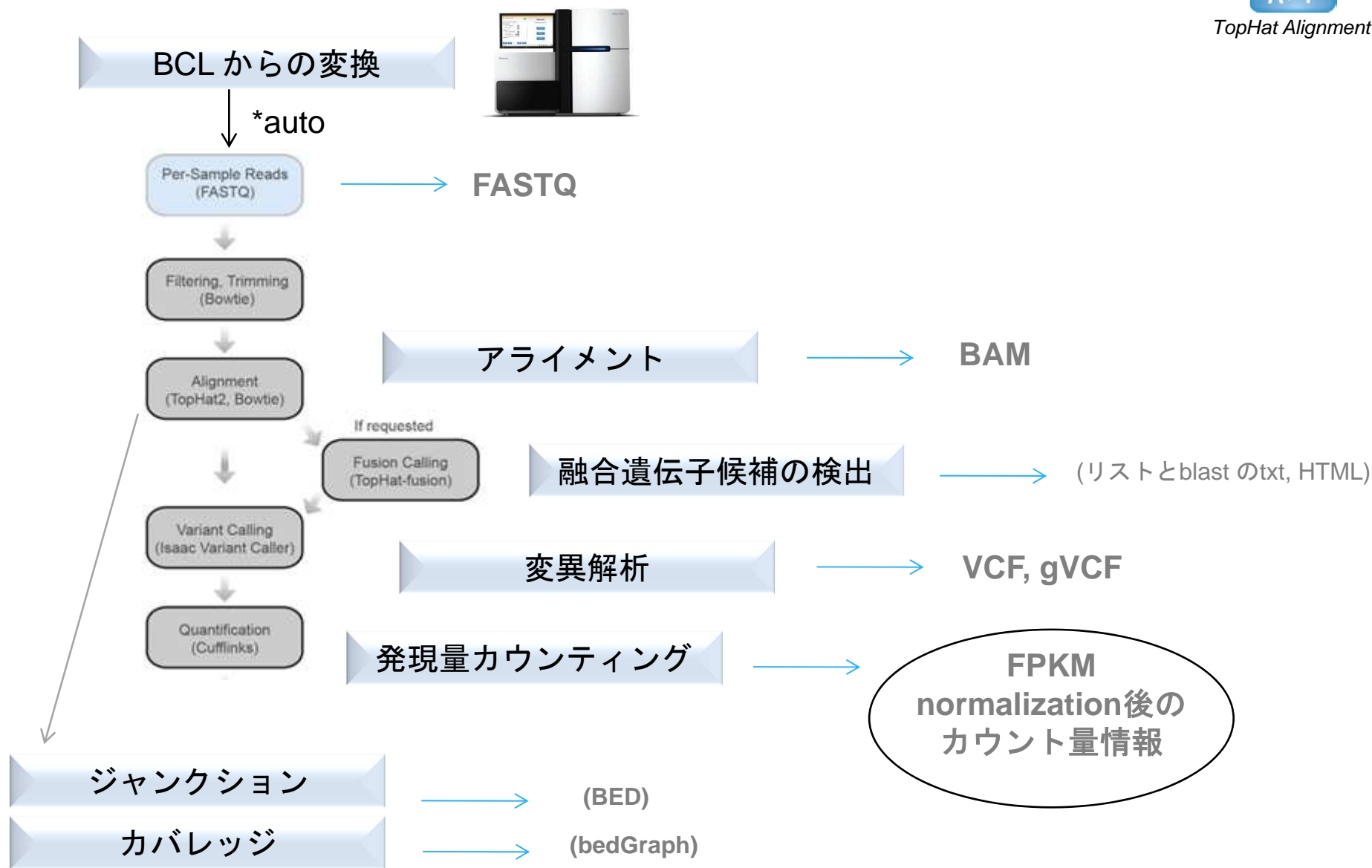


TopHat 本家マニュアル :  
<http://tophat.cbcb.umd.edu/manual.shtml>

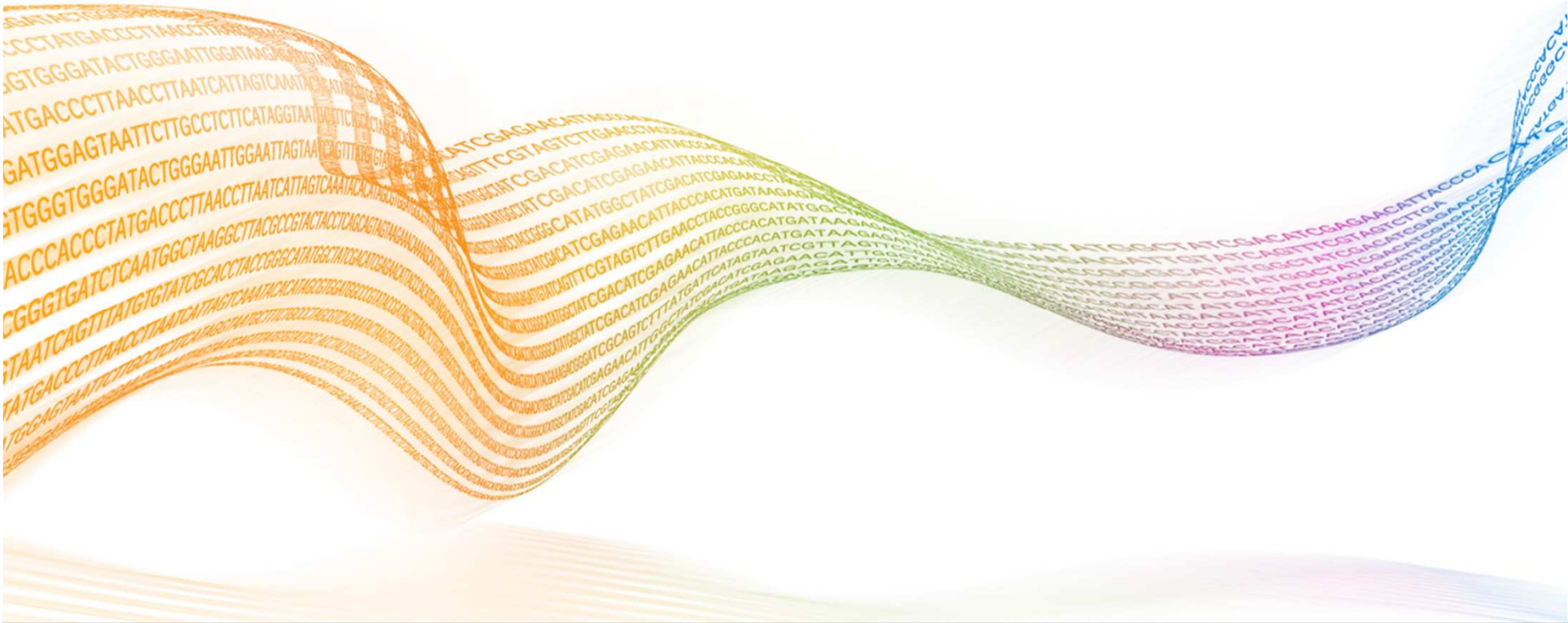
# TopHat Alignment App のパイプライン各工程と出力ファイル形式



TopHat Alignment



# BaseSpace Cufflinks & DE アプリ

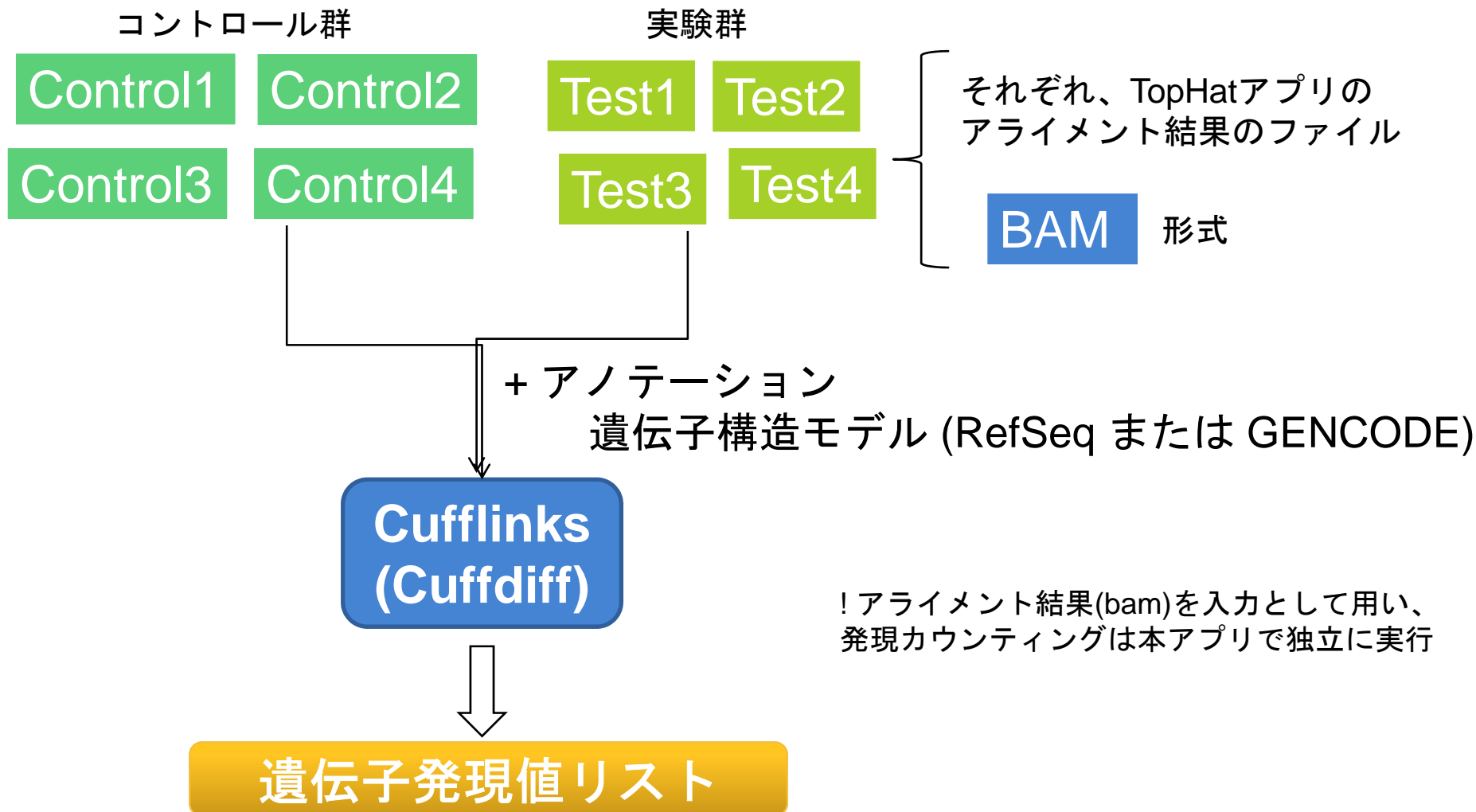


# Cufflinks によるDEの概要



Cufflinks Assembly/  
Diff. Exp

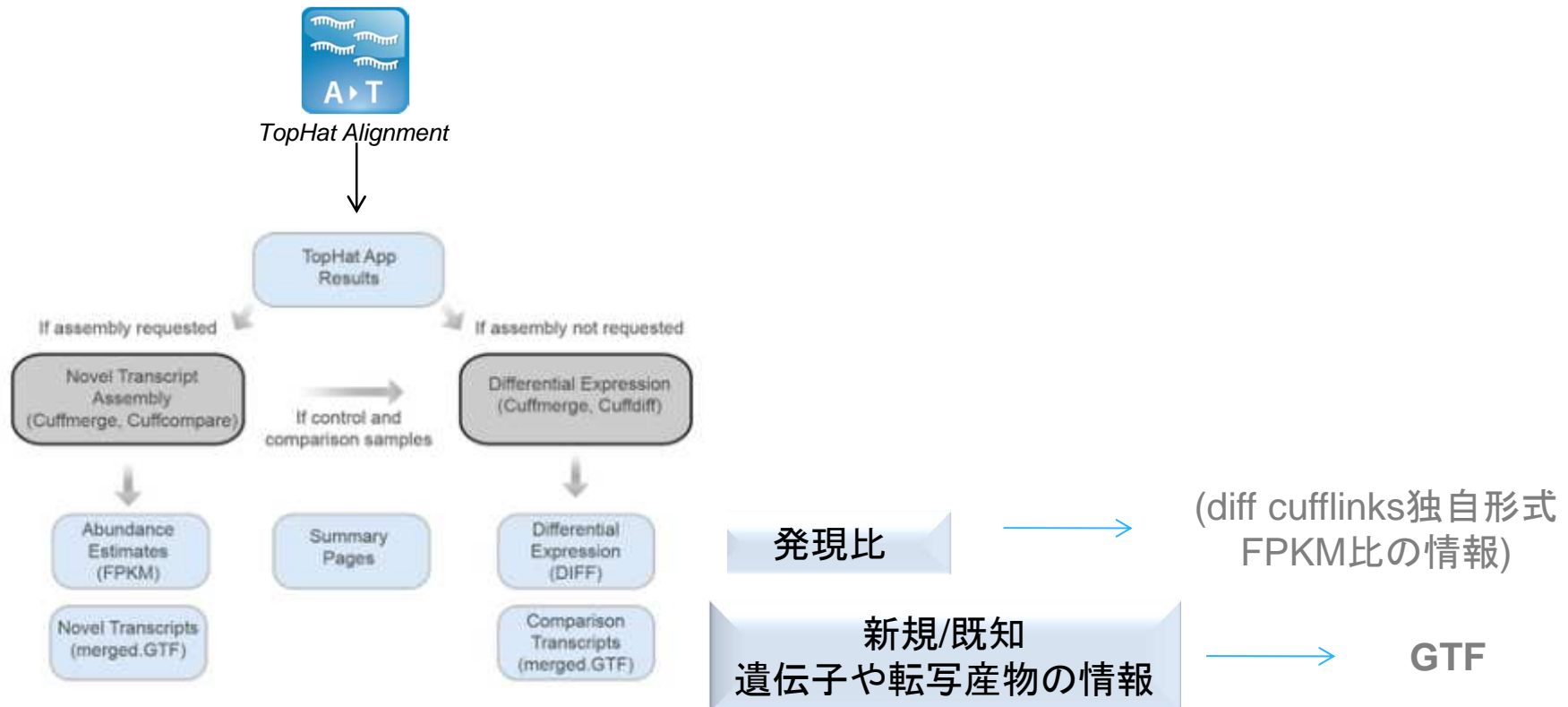
バイオロジカルレプリケート推奨 (以下の例ではレプリケート数 n=4)



# Cufflinks Assembly & DE App のパイプライン



Cufflinks Assembly/  
Diff. Exp



ユーザガイドに記載されていない詳細は、  
Cufflinks 本家マニュアルをご参考いただけます <http://cufflinks.cbcb.umd.edu/manual.html>

# Cufflink Assembly & DE 設定画面



App Session Name: Cufflinks Assembly & DE 05/16/2

Save Results To: Select Project(s):  
Transcriptome Demo

## TopHat Alignments Selection Criteria

Reference Genome: Homo sapiens/hg19 (RefSeq)

Stranded:  Stranded キットで調製を行った場合

## Options

Novel Transcript Assembly:  新規転写産物のアセンブルを行う場合

### 参照データベースの選択

Homo sapiens/hg19 (RefSeq)  
**Homo sapiens/hg19 (RefSeq)**  
Homo sapiens/hg19 (Gencode)  
Mus musculus/mm10 (RefSeq)  
Rattus norvegicus/rn5 (RefSeq)

! TopHatアプリ実行時の  
選択と整合している  
必要があります。



# Cufflink Assembly & DE 設定画面



## Control Group

Group Label: control

TopHat Alignment App Result(s):

Select App Result(s):

mRNA-UHRR-C2	x
mRNA-UHRR-C1	x

Adjust transcript assembly for samples without polyA selection:

## Comparison Group

Group Label: comparison

TopHat Alignment App Result(s):

Select App Result(s):

mRNA-Brain-C4	x
mRNA-Brain-C6	x

Adjust transcript assembly for samples without polyA selection:

**Control Group**  
TopHatアプリの結果を選択する

**Comparison Group**  
TopHat アプリの結果を選択する

i) TruSeq Stranded Total RNA KitのようなPolyA選別していない場合にチェックを入れ最適化。

# Cufflinks アプリ出力結果例



BaseSpace

HiSeq 2000: TruSeq Stranded Total RNA (MAQC) : Cufflinks Assembly & DE 04/24/2014 6:18:02

Analysis Info

Inputs

Output Files

Analysis Reports

Cufflinks-Report

### Overview

Control samples (UHR)

- RZ100ngUHR-i5-A1-01
- RZ100ngUHR-i5-B1-02
- RZ100ngUHR-i5-C1-03
- RZ100ngUHR-i5-D1-04

Comparison samples (Brain)

- RZ100ngHuBr-i6-E1-01
- RZ100ngHuBr-i6-F1-02
- RZ100ngHuBr-i6-G1-03
- RZ100ngHuBr-i6-H1-04

FPKM tables: Genes / Transcripts

# Cufflink Assembly & DE 実行結果



Assembly <sup>i</sup>

	Control	Comparison	Merged
Gene Count	45,635	52,884	62,231
Transcript Count	87,392	96,262	119,726
Link to gene models	<a href="#">GTF result</a>	<a href="#">GTF result</a>	<a href="#">GTF result</a>
Relation to reference transcripts			
Equal (=)	44,934	45,437	46,093
Potentially novel (l)	18,903	20,155	31,453
Unknown, Intergenic (u)	21,848	28,622	39,115
Overlap with opposite-strand exon (x)	1,361	1,674	2,459
Other	346	374	606

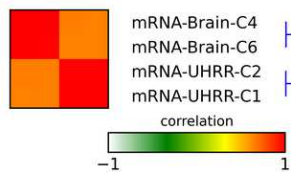
遺伝子・トランスクリプトのカウント数や遺伝子構造について

Differential Expression <sup>i</sup>

Gene Count	62,225
$\Delta$ Gene Count	27,052
Transcript Count	119,696
$\Delta$ Transcript Count	27,379
CuffDiff results	<a href="#">differential gene expression, differential transcript expression</a>

発現量の差異が有意にみられた遺伝子・トランスクリプトの数

## Sample Correlation

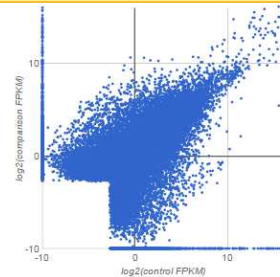


サンプル間のcorrelation

## Differential Expression Gene Browser



DE Gene Browser の  
スキャタープロット



発現差のある遺伝子

Significant

false  
true

特定遺伝子の表示

Gene

# Cufflinksアプリ出力結果例 (各サンプル毎のFPKMリスト)



レポート中に表示されるリンク→

FPKM tables: [Genes](#) / [Transcripts](#)

コントロール群

実験群

A	B	C	D	E	F	G	H	I	J
		RZ100ngHuBr- i6-E1-	RZ100ngHuBr- i6-F1-	RZ100ngHuBr- i6-G1-	RZ100ngHuBr- i6-H1-	RZ100ngUHR- i5-A1-	RZ100ngUHR- i5-B1-	RZ100ngUHR- i5-C1-	RZ100ngUHR- i5-D1-
tracking id	locus	01.FPKM	02.FPKM	03.FPKM	04.FPKM	01.FPKM	02.FPKM	03.FPKM	04.FPKM
A1BG	chr19:588	0.877336	0.751462	0.791959	0.585528	1.50524	1.53949	1.55188	1.47948
A1BG-AS1	chr19:588	0.55513	0.600267	0.613914	0.759932	0.917981	0.725987	0.650588	0.911839
A1CF	chr10:525	0.0290704	0.0371743	0.0366497	0.0397232	1.5566	1.41757	1.47525	1.46501
A2M	chr12:922	20.9978	21.0211	21.3772	21.8963	69.719	69.1105	68.7974	68.2234
A2M-AS1	chr12:921	1.109	1.11787	1.14053	1.20411	0.423059	0.492536	0.558136	0.625029
A2ML1	chr12:897	0.588004	0.647205	0.647019	0.654925	0.227352	0.22362	0.177014	0.177929
A2MP1	chr12:938	0.102718	0.0951737	0.104212	0.0784438	0.0210817	0.0623826	0.0492729	0.0494728
A4GALT	chr22:430	0.572499	0.428008	0.602284	0.502222	1.20627	1.53882	1.30407	1.41978
A4GNT	chr3:1378	0.001	0.0102621	0.0192562	0.00845964	0.0272833	0.0673127	0.001	0.001
AA06	chr17:318	0.001	0.418099	0.001	0.200904	0.001	0.001	0.001	0.001
AAAS	chr12:537	4.76377	4.74959	5.36134	5.06117	16.708	15.4295	16.1457	17.2751
AACS	chr12:125	7.08618	6.89536	6.74231	6.57433	5.36705	5.31277	5.37381	5.52981
AACSP1	chr5:1781	0.0974901	0.0821632	0.115616	0.0609623	0.425995	0.62533	0.421505	0.411009
AADAC	chr3:1515	0.001	0.0114303	0.001	0.001	0.0303877	0.0449782	0.0532871	0.0356999
AADACL2	chr3:1514	0.001	0.001	0.001	0.001	0.016392	0.001	0.0191617	0.001
AADACL3	chr1:1277	0.001	0.001	0.001	0.001	0.0282973	0.0236709	0.0210312	0.00705382
AADACL4	chr1:1270	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
AADAT	chr4:1709	1.51119	1.79033	1.64225	1.58242	1.61109	1.81217	1.70279	1.65033
AAED1	chr9:9940	1.55632	1.95375	1.96462	1.78926	6.06918	5.9358	6.16553	6.31189

レプリケート4ずつ

# Cufflinksアプリ出力 - 遺伝子発現差異リスト



コントロール群                      実験群                      2群間の差異

Filters  
log2(ratio) 0.0 4.0

	A	B	C	D	E	F	G	H	I
	Test ID	Gene	Locus	Status	log2(contr	log2(com	log2(Ratio	Value	Significant
2	A1CF	A1CF	chr10:525	OK	-3.86	1.41	-5.27	6.33E-05	TRUE
3	A2ML1	A2ML1	chr12:897	OK	0.24	-2.41	2.65	6.33E-05	TRUE
4	AACSP1	AACSP1	chr5:1781	OK	-2.33	-0.24	-2.1	6.33E-05	TRUE
5	AAK1	AAK1	chr2:6968	OK	3.98	1.63	2.35	6.33E-05	TRUE
6	AATK	AATK	chr17:790	OK	4.99	0.29	4.7	6.33E-05	TRUE
7	AATK-AS1	AATK-AS1	chr17:790	OK	-1.14	-3.72	2.58	0.003908	TRUE
8	ABAT	ABAT	chr16:876	OK	6.29	2.64	3.65	6.33E-05	TRUE
9	ABCA10	ABCA10	chr17:671	OK	1.78	-2.55	4.32	6.33E-05	TRUE
10	ABCA12	ABCA12	chr2:2157	OK	-4.37	-0.49	-3.88	6.33E-05	TRUE
11	ABCA2	ABCA2	chr9:1399	OK	6.45	3.55	2.9	6.33E-05	TRUE
12	ABCA3	ABCA3	chr16:232	OK	4.98	2.5	2.48	6.33E-05	TRUE
13	ABCA5	ABCA5	chr17:671	OK	4.55	2.51	2.04	6.33E-05	TRUE
14	ABCA6	ABCA6	chr17:670	OK	2.39	-4.04	6.43	6.33E-05	TRUE
15	ABCA8	ABCA8	chr17:668	OK	3.89	0.04	3.86	6.33E-05	TRUE
16	ABCA9	ABCA9	chr17:669	OK	2.15	-0.51	2.66	6.33E-05	TRUE
17	ABCB5	ABCB5	chr7:2065	OK	-0.65	1.91	-2.56	6.33E-05	TRUE

# Cufflinksアプリ DEフィルタリング



## Differential Expression Gene Browser

### Filters

**||log<sub>2</sub>(ratio)|**  
 2.0  23.0

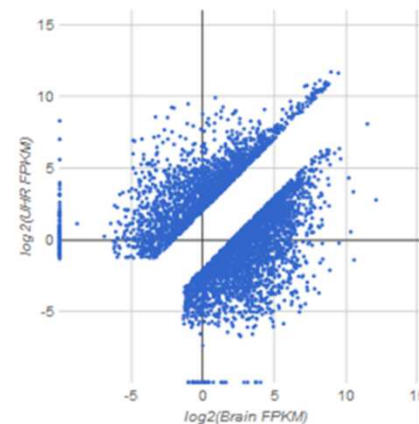
### Significant

true ▾

### Status

OK ▾

### Gene



Save Plot as SVG

### Differential Expression Gene Browser

The Differential Expression Gene Browser can be used to filter and plot the differential expression results dynamically (Table 5). The data can be sorted by clicking any column heading and can be saved as either an SVG graphic or CSV table.

Table 5: Differential Expression Gene Browser Filter Options

Filter	Description
Log Ratio Cutoff	Allows the differentially expressed gene table to be filtered based on the observed log ratio of differential expression.
Significance	Filters for results achieving statistical significance at a q-value < 0.05.
Status	Filters based on the reliability of the result. If the result passes all stringency filters, the status is returned as "OK". If the result is flagged due to one of several factors (such as insufficient read depth), a failure mode of NOTEST, LOWDATA, or FAIL is returned.

Additional information is available at <http://cufflinks.cbcb.umd.edu/faq.html#notest>.

Test ID	Gene	Locus	Status	log <sub>2</sub> (Brain FPKM)	log <sub>2</sub> (UHR FPKM)	log <sub>2</sub> (Ratio)	q Value	Significant
A1CF	A1CF	chr10:52559168-52645435	OK	-3.860	1.410	-5.270	0.000	✓
A2ML1	A2ML1	chr12:89:					0.000	✓
AACSP1	AACSP1	chr5:178:					0.000	✓
AAK1	AAK1	chr2:698:					0.000	✓

Save Filtered Table

テクニカルノートより

[www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote-basespace-rna-seq.pdf](http://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote-basespace-rna-seq.pdf)

ユーザガイドにも詳細がございます

# Cufflink Assembly & DE 主要出力ファイルまとめ



## FPKM File

検体ごとの遺伝子とアイソフォームの発現量 (FPKM)を示す

#	A	B	C	D	E	F	G	H	I	J	K	L	M
1	tracking_id	class_code	nearest_ref	gene_id	gene_short	tss_id	locus	length	coverage	FPKM	FPKM_conf	FPKM_conf	FPKM_status
2	TCONS_00=		NR_04601	EXLOC_0001	DDX11L1	TSS1	chr1:11875	1652	0.272117	0.034342	0.006323	0.066588	OK
3	TCONS_00=		NR_02454	XLOC_0021	WASH7P	TSS2909	chr1:14361	1769	53.0076	6.6898	6.1494	7.23408	OK
4	TCONS_00=		NR_02682	XLOC_0021	FAM138F	TSS2910	chr1:3461C	1130	0.186993	0.017702	0	0.060842	OK
5	TCONS_00=		NM_00100	XLOC_0001	OR4F5	TSS2	chr1:6909C	918	0	0	0	0	OK
6	TCONS_00=		NR_03998	XLOC_0021	LOC72973	TSS2911	chr1:13477	5474	83.2021	14.7687	14.0317	15.1068	OK
7	TCONS_00=		NR_02832	XLOC_0001	LOC10013	TSS3	chr1:32398	4370	15.2952	2.79301	2.51434	3.07599	OK
8	TCONS_00=		NR_02832	XLOC_0001	LOC10013	TSS3	chr1:32398	4273	69.9856	12.6173	12.21529	13.2492	OK
9	TCONS_00=		NM_00100	XLOC_0001	OR4F3	TSS4	chr1:3676E	939	0.464128	0.068148	0.029287	0.16108	OK
10	TCONS_00=			XLOC_0001		TSS5	chr1:5671E	81	9590.67	2214.32	1.69758	5.26249	OK
11	TCONS_00=		NM_00100	XLOC_0021	OR4F3	TSS2912	chr1:6210E	939	0.464128	0.068148	0.029287	0.16108	OK
12	TCONS_00=		NR_02832	XLOC_0021	LOC10013	TSS2913	chr1:6611E	4273	80.5244	14.7087	13.8115	14.9475	OK
13	TCONS_00j		NR_03990	XLOC_0021	LOC10028	TSS2914	chr1:69421	1863	11.8876	1.87101	1.55745	2.18688	OK
14	TCONS_00=		NR_03990	XLOC_0021	LOC10028	TSS2915	chr1:70024	1371	72.1464	11.3553	10.4833	12.2305	OK
15	TCONS_00=			XLOC_0001		TSS6	chr1:71442	1082	3.82708	0.59841	0.381249	0.813331	OK
16	TCONS_00=			XLOC_0001		TSS7	chr1:7173E	5853	7.04468	1.1731	1.05483	1.28976	OK
17	TCONS_00=			XLOC_0001		TSS8	chr1:7255E	600	0.489808	0.093542	0.022917	0.183338	OK
18	TCONS_00=			XLOC_0001		TSS9	chr1:72657	1253	0.437604	0.069132	0.010974	0.120714	OK
19	TCONS_00=			XLOC_0021		TSS2916	chr1:72882	1332	2.71018	0.404437	0.258078	0.547124	OK
20	TCONS_00=			XLOC_0021		TSS2917	chr1:73022	1192	0.781542	0.179707	0.092284	0.265318	OK
21	TCONS_00=			XLOC_0001		TSS10	chr1:7321E	1004	1.43742	0.227828	0.109656	0.34239	OK
22	TCONS_00=			XLOC_0001		TSS2918	chr1:7326E	839	0.206266	0.326929	0.161779	0.485338	OK
23	TCONS_00=			XLOC_0001		TSS11	chr1:7372C	1233	0.374999	0.18565	0.089216	0.278799	OK
24	TCONS_00=			XLOC_0001		TSS12	chr1:7391E	648	1.22792	0.231999	0.084879	0.360735	OK
25	TCONS_00=			XLOC_0021		TSS2919	chr1:74834	780	5.34747	0.819321	0.475974	1.09298	OK
26	TCONS_00=			XLOC_0021		TSS2920	chr1:7525E	1000	3.36759	0.343368	0.192505	0.481263	OK
27	TCONS_00=		NR_02432	XLOC_0021	LINC0011E	TSS2921	chr1:7615E	1317	23.9665	2.47587	1.93153	2.7459	OK
28	TCONS_00j		NR_01536	XLOC_0001	LOC64383	TSS13	chr1:7629E	6385	2.15434	0.219462	0.15239	0.287112	OK

## DIFF File

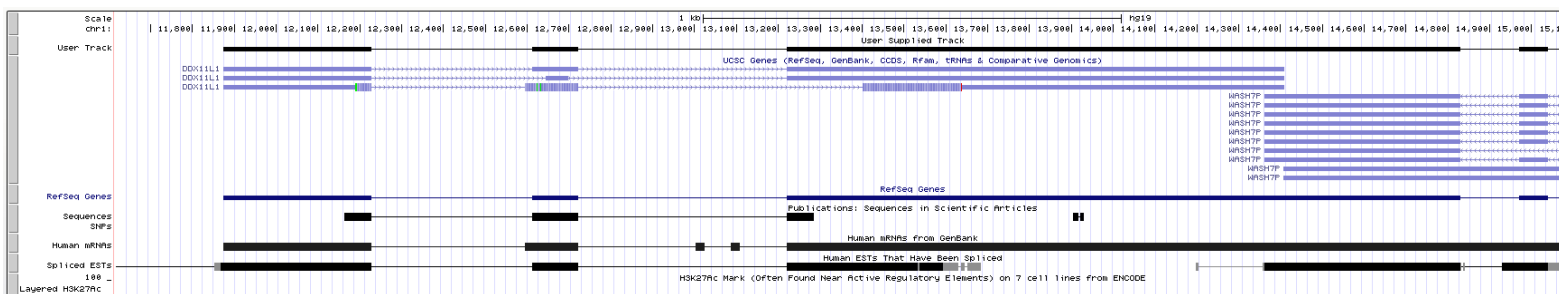
Control GroupとComparison Groupとの発現量の比較

#	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	test_id	gene_id	locus	sample_1	sample_2	status	value_1	value_2	log2(fold)	ci	test_stat	p_value	q_value	significant	
2	TCONS_00	XLOC_0001	DDX11L1	chr1:11875	control	comparator	NOT EST	0	0	0	0	0	1	no	
3	TCONS_00	XLOC_0001	OR4F5	chr1:6909C	control	comparator	NOT EST	0	0	0	0	0	1	no	
4	TCONS_00	XLOC_0001	LOC10013	chr1:32398	control	comparator	NOT EST	3.29E-05	0.004898	7.21688	0	0	1	no	
5	TCONS_00	XLOC_0001	LOC10013	chr1:32398	control	comparator	NOT EST	0.001838	0	#NAME?	0	0	1	no	
6	TCONS_00	XLOC_0001	OR4F3	chr1:3676E	control	comparator	NOT EST	0	0	0	0	0	1	no	
7	TCONS_00	XLOC_0001		chr1:5671E	control	comparator	OK	0	7524.08	inf	#NAME?	0.00015	0.001006	yes	
8	TCONS_00	XLOC_0001		chr1:56711	control	comparator	OK	32.3687	33.3108	0.041395	#NAME?	0.734084	0.96835	0.974063	no
9	TCONS_00	XLOC_0001		chr1:56781	control	comparator	OK	2.13267	4.19778	0.976964	2.13267	4.19778	0.33815	0.437867	no
10	TCONS_00	XLOC_0001		chr1:71442	control	comparator	OK	0.380312	0.285006	-0.41619	-0.6112	0.43345	0.525401	no	
11	TCONS_00	XLOC_0001		chr1:7173E	control	comparator	OK	0.690636	0.456799	-0.59637	-0.99882	0.1692	0.319522	no	
12	TCONS_00	XLOC_0001		chr1:7173E	control	comparator	OK	0.548731	0.980368	0.837225	0.714441	0.3312	0.431247	no	
13	TCONS_00	XLOC_0001		chr1:7255E	control	comparator	OK	0.262932	0	#NAME?	#NAME?	0.00075	0.004225	yes	
14	TCONS_00	XLOC_0001		chr1:72657	control	comparator	NOT EST	0.049307	0.012572	-1.9716	0	1	1	no	
15	TCONS_00	XLOC_0001		chr1:7321E	control	comparator	NOT EST	0.100724	0.025186	-1.99972	0	1	1	no	
16	TCONS_00	XLOC_0001		chr1:7372C	control	comparator	NOT EST	0.12213	0.066124	-0.88517	0	1	1	no	
17	TCONS_00	XLOC_0001		chr1:7391E	control	comparator	OK	0.316741	0.34727	0.132754	0.1279	0.8932	0.916355	no	
18	TCONS_00	XLOC_0001		chr1:7543E	control	comparator	OK	0.408285	0.613477	0.587434	0.600664	0.62025	0.691082	no	
19	TCONS_00	XLOC_0001		chr1:7548E	control	comparator	OK	0.10476	0.208833	1.00215	0.636658	0.50615	0.590781	no	
20	TCONS_00	XLOC_0001		chr1:7599E	control	comparator	OK	0.074714	0.371156	2.31257	3.06832	0.0024	0.011357	yes	
21	TCONS_00	XLOC_0001	LOC64383	chr1:75971	control	comparator	OK	0	1.99597	inf	#NAME?	0.1257	0.271095	no	
22	TCONS_00	XLOC_0001	LOC64383	chr1:75971	control	comparator	NOT EST	0	0	0	0	0	1	no	
23	TCONS_00	XLOC_0001	LOC64383	chr1:75977	control	comparator	OK	0	0.71728	inf	#NAME?	0.1239	0.271095	no	
24	TCONS_00	XLOC_0001	LOC64383	chr1:75977	control	comparator	OK	0.778959	0.025581	-4.92841	-1.50233	0.1939	0.3338	no	
25	TCONS_00	XLOC_0001	LOC64383	chr1:75977	control	comparator	OK	1.30685	0.463559	-1.49527	-1.3108	0.09005	0.235133	no	
26	TCONS_00	XLOC_0001	LOC64383	chr1:75977	control	comparator	OK	0	0.487844	inf	#NAME?	0.12395	0.271095	no	
27	TCONS_00	XLOC_0001	LOC64383	chr1:75977	control	comparator	NOT EST	0.030163	0.104867	1.7977	0	1	1	no	
28	TCONS_00	XLOC_0001	LOC64383	chr1:75977	control	comparator	OK	0.448803	2.77794	2.62986	3.08724	0.0013	0.006788	yes	
29	TCONS_00	XLOC_0001	LOC64383	chr1:75977	control	comparator	OK	0	3.07596	inf	#NAME?	0.1240	0.271095	no	

## GTF file

転写産物の構造を示すファイル

UCSC Genome Browserを用いた表示例



追加された  
GTFファイル  
のトラック

# ステップ・バイ・ステップの説明 デモ動画



[www.illumina.com/informatics/research/sequencing-data-analysis-management/rna-seq-data-analysis.html](http://www.illumina.com/informatics/research/sequencing-data-analysis-management/rna-seq-data-analysis.html)

**RNA-Seq BaseSpace Apps: A Guided Tour**

See step-by-step instructions on how to navigate through the data analysis.

[View Video](#)

## RNA-Seq BaseSpace Apps: A Guided Tour

Test ID	Gene	Locus	Status	log <sub>2</sub> (uhr FPKM)	log <sub>2</sub> (brain FPKM)	log <sub>2</sub> (Ratio)	q Value	Significant
XLOC_000010	-	chr1:797247-799101	OK	-10.000	-0.740	-9.260	0.000	✓
XLOC_000011	-	chr1:800366-801218	OK	-10.000	-0.630	-9.370	0.000	✓
XLOC_000014	-	chr1:844861-845337	OK	-0.540	-10.000	9.460	0.001	✓
XLOC_000028	-	chr1:1100832-1101478	OK	-0.920	-10.000	9.080	0.000	✓
XLOC_000043	-	chr1:1314123-1314431	OK	0.040	-10.000	10.040	0.003	✓
XLOC_000062	GABRD	chr1:1950767-1962192	OK	-1.350	6.950	-8.300	0.000	✓
XLOC_000068	PLCH2	chr1:2398901-2439211	OK	-0.690	4.460	-5.150	0.000	✓
XLOC_000069	-	chr1:2462986-2463443	OK	-10.000	-0.700	-9.300	0.002	✓
XLOC_000070	-	chr1:2469531-2469979	OK	-10.000	-0.870	-9.130	0.002	✓
XLOC_000071	-	chr1:2470381-2470659	OK	-10.000	0.630	-10.630	0.001	✓
XLOC_000072	-	chr1:2472405-2473064	OK	-10.000	-0.980	-9.020	0.000	✓
XLOC_000073	-	chr1:2475563-2477232	OK	-10.000	-0.540	-9.460	0.000	✓
XLOC_000078	-	chr1:2495662-2500317	OK	-0.370	-10.000	9.630	0.000	✓
XLOC_000085	-	chr1:4036939-	OK	3.320	-4.210	-7.530	0.030	✓



## さらなる解析のために

解析が具体的にどのように実行されたかは、AnalysisInfoの  
Log Filesに記載あり

The screenshot shows the BaseSpace interface for a 'TopHat Alignment' job. The 'Analysis Info' tab is selected and highlighted with a red circle. The job details are as follows:

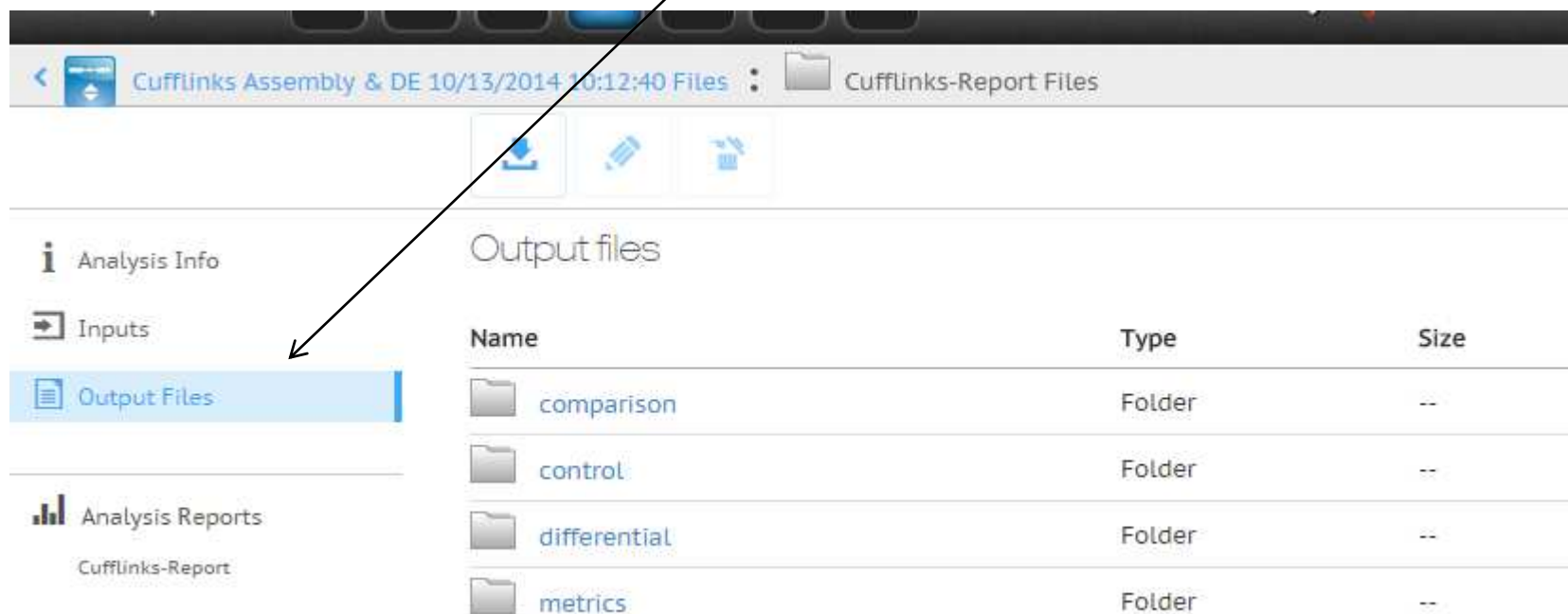
Name	TopHat Alignment 10/07/2014 1:57:42
Application	TopHat Alignment   Version: 1.0.0
Date started	Tuesday, October 7th 2014, 8:58:04 pm
Date completed	Wednesday, October 8th 2014, 7:15:18 pm
Duration	22 hours 17 minutes 14 seconds
Session Type	Multi-Node
Size	81.95 GB
Status	Complete (8 Nodes Complete)

At the bottom of the page, there is a 'Log Files' link and a note: 'Please view the Multi-Node details page to see logs for this analysis.'

！ BaseSpaceのインターフェイスにおいては、コマンドラインの利用と比較して、簡単に最適なオプションを素早く選択できるように、変更可能な解析条件は絞られ予めプリフィックスされ設計されている。

！ しかし実際の使用オプションや詳細な工程の順を把握したい場合は、ログファイルを追うことである程度これを把握できる。

## さらなる解析のために 解析結果の生ファイルは **Output Files**の配下にあります

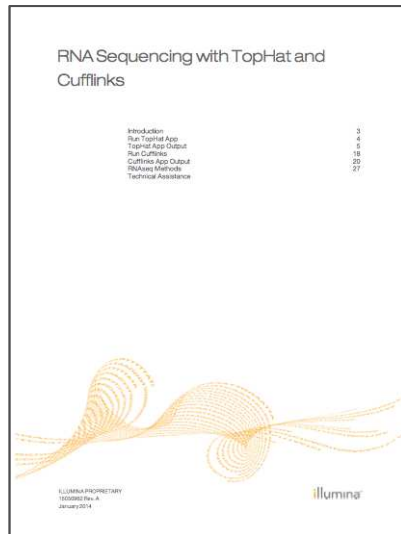


! 全てのoutput項目がpdfレポート表示されているわけではない。  
isoform毎の発現リスト等 実行結果の詳細は全てOutputFiles配下に置かれている。

! R/BioConductorなど他のツールで更に解析をすすめたり可視化を行う場合は  
このOutput Filesの中から様々なフォーマットのファイルをダウンロードして利用できる。



# 詳細は User Guide もご参考ください

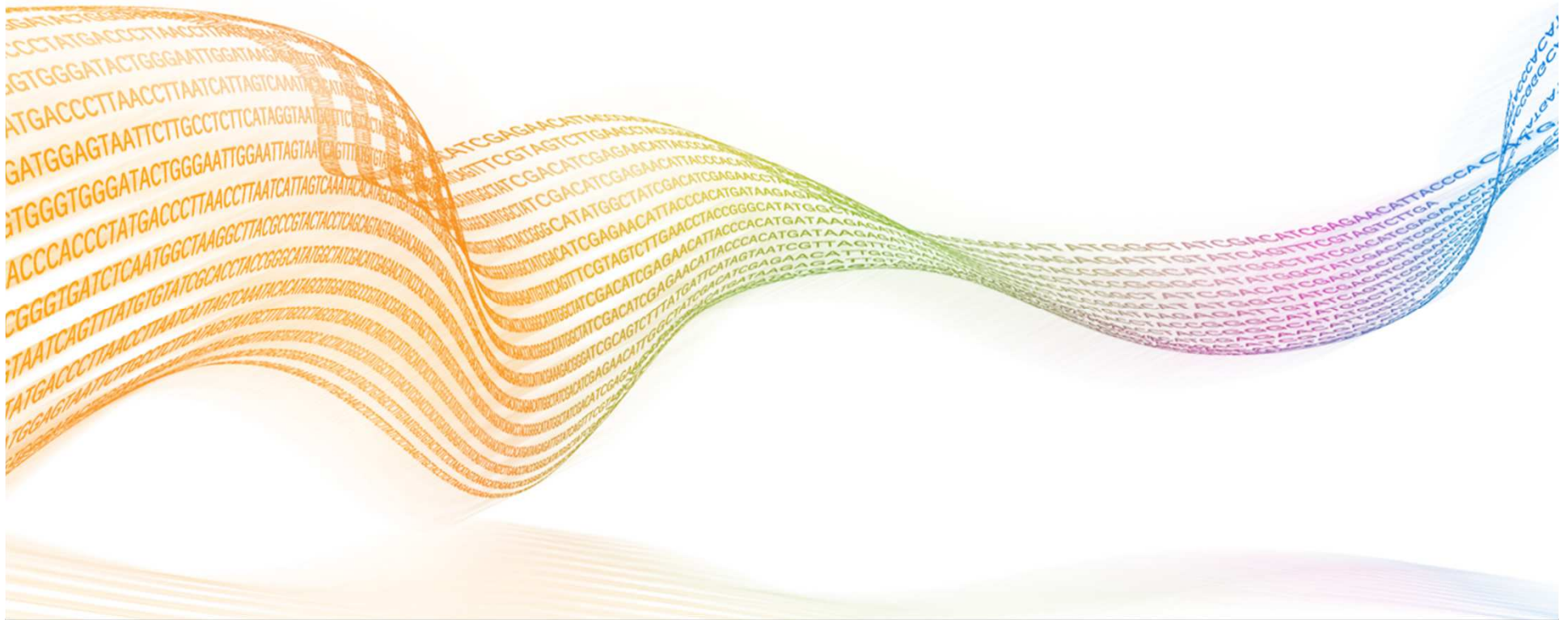


RNA-Seq  
TopHat  
Cufflinks



RNA  
Express

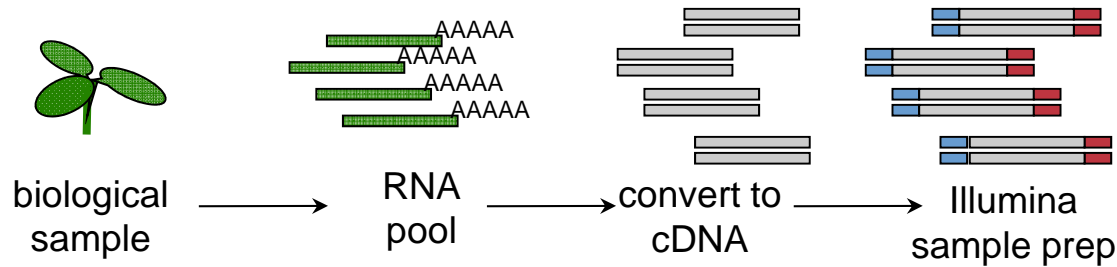
# シーケンシング以前： 実験デザインの解析結果への影響



# RNA-Seq シーケンス前

どんな RNA を調べたい？  
mRNA？  
トータルRNA？

発現量を数えるのみ？  
新規転写産物情報まで解析  
したいか



サンプルの知見  
劣化状況は？  
準備可能なRNA量は？

ストランド情報を  
保持したい？

カバレッジは？  
PE/SR？  
シーケンス長？

# RNA Seq の投入リード数の大まかな目安

mRNA ***	Differential Expression	10-20M Reads *
	Allele Specific Expression	50 – 100M Reads *
	Splice variation **	50 - 100M Reads *
	Complete Annotation	100M - 1B Reads *
	Transcript Based Assembly	50 – 200M Reads *


76bp~ x PE


- Based on human sized transcriptomes
- \*\* Also applies to RNA fusion transcripts in cancer
- \*\*\* Applies to poly A-selected libraries
  - Ribo-Zero, high quality RNA libraries: requires ~ 2X more reads
  - Ribo-Zero, FFPE libraries: requires ~ 4X more reads

- あくまで目安となります。
- 業界ガイドラインご参考 **ENCODE** ;  
<https://www.encodeproject.org/about/experiment-guidelines/#guideline>

# 解析ツール側での入力条件との兼ね合い

BaseSpace  
TopHat Alignmentの例

 **TopHat Alignment**  
Illumina, Inc. [Website »](#)

version 1.0.0  [Launch](#)

**Free**

CATEGORIES RNA-Seq, Gene Fusion Detection, Tumor Normal

**License Info**  
[Privacy](#)  
[EULA](#)

**Description**

The TopHat Alignment workflow performs the following functions

- Read mapping using the TopHat 2 aligner
- FPKM estimation of reference genes and transcripts using Cufflinks 2
- Variant calling (SNVs and small indels) with the Isaac Variant caller
- Optional fusion calling with TopHat-Fusion

Available reference genomes include

- Homo sapiens UCSC hg19 (RefSeq & Gencode gene annotations)
- Mus musculus UCSC mm10 (RefSeq gene annotations)
- Rattus norvegicus UCSC m5 (RefSeq gene annotations)

Current limitations:

- Minimum and maximum of 100,000 and 400,000,000 reads per-sample, respectively
- Maximum of 2,000,000,000 read across all samples in a single analysis
- Minimum Read length 35bp and Maximum Read Length 125bp
- Paired-end reads are required for fusion detection



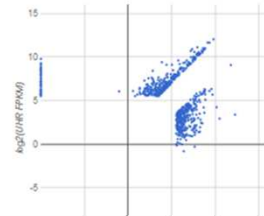
# DEにおけるカバレッジの影響

500K reads/sample

Differential Expression Gene



740 diff genes

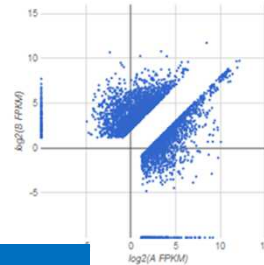


10 million reads/sample

Differential Expression Gene

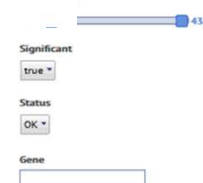


3,800 diff genes

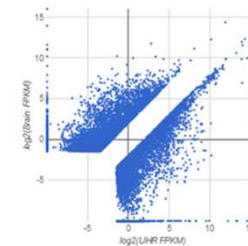


60 million reads/sample

Differential Expression Gene



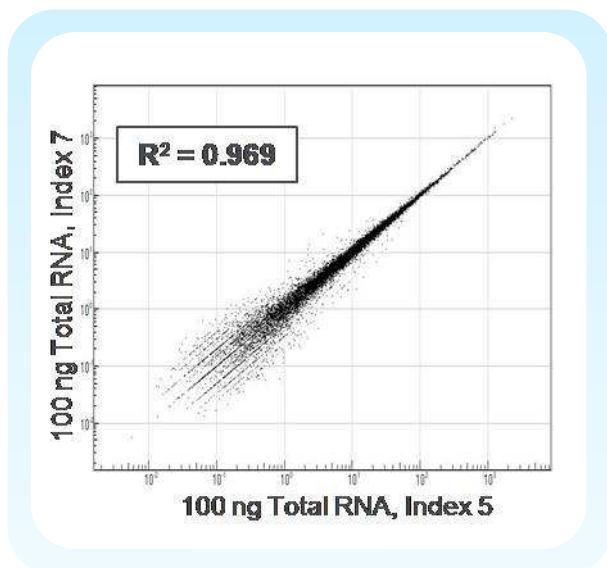
5,400 diff genes



BaseSpaceのCufflinks & DE assemblyアプリを利用

# レプリケーションの検討

High Reproducibility  
with Low Inputs



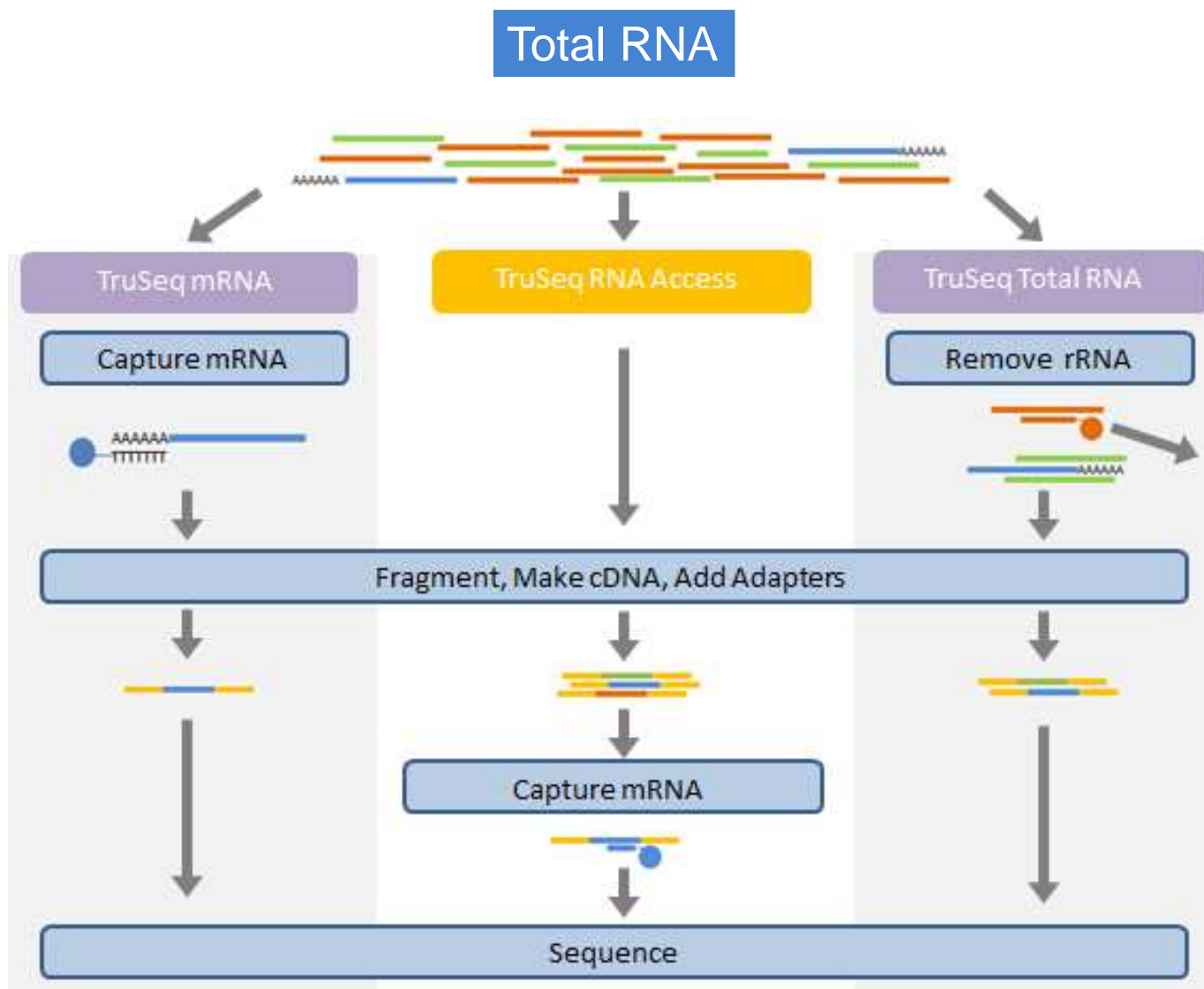
- ▶ テクニカルレプリケーションは  
イルミナシーケンサーではほとんど  
必用ない
- ▶ バイオロジカルレプリケーション  
は推奨される

Correlation Coefficient 0.92-0.98

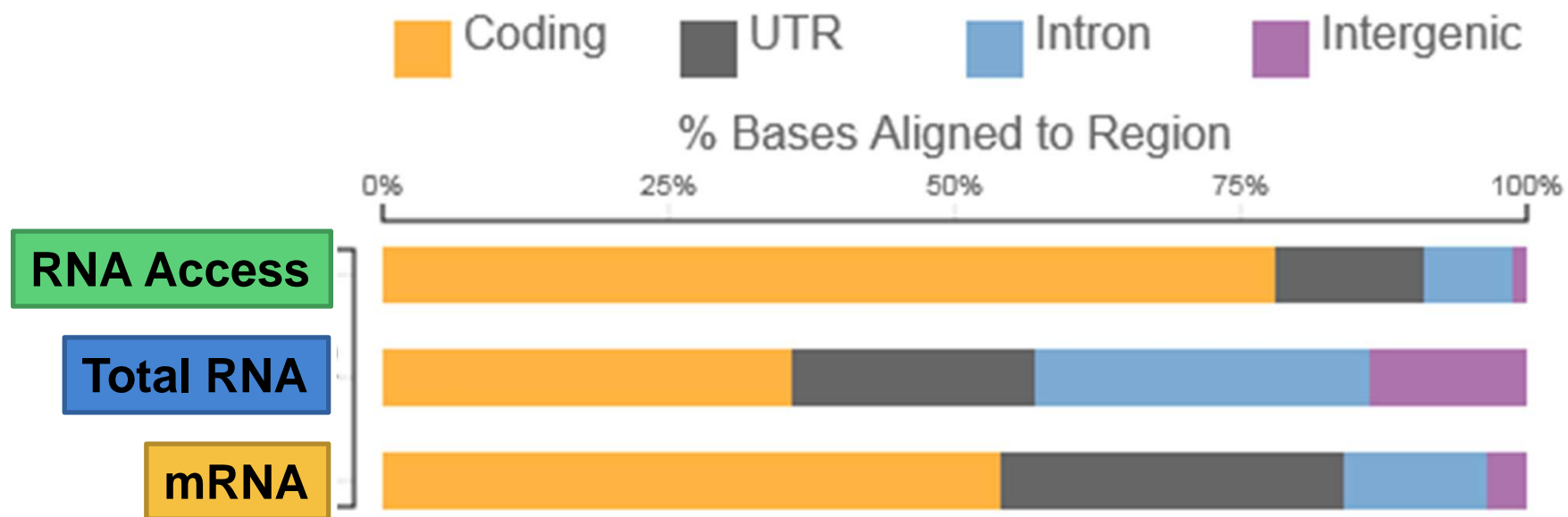
実験設計のガイドライン/標準/ベストプラクティス；

<https://www.encodeproject.org/about/experiment-guidelines/#guideline>

# キットの特性からの検討



# キット毎のリード分布



※ BaseSpaceのTopHatアプリのレポートからQC結果を抜粋

# キット別の遺伝子上のマップ例



# Upcoming Webinar

- ▶ **"Next generation tools for gene expression profiling and gene fusion detection in FFPE tumor samples"**
  - Gary P. Schroth Ph.D., Distinguished Scientist, Illumina
- ▶ **10月 13日 (火) 13:00~**
- ▶ **Register NOW!**
  - <https://illumina.webex.com/illumina/onstage/g.php?2f9bc840e5ddbfc22fcb31d25>



# ご参考文献

## ▶ イルミナシーケンシング

- Accurate whole human genome sequencing using reversible terminator chemistry.  
Nature 456: 53-59 [PMID: 18987734]

## ▶ RNA-Seq, RPKM

- Mapping and quantifying mammalian transcriptomes by RNA-Seq  
Nature Methods, Volume 5, 621 – 628 [PMID: 18516045]

## ▶ BaseSpace RNA-Seq アプリ

- TopHat: discovering splice junctions with RNA-Seq.  
Bioinformatics. 2009 May 1;25(9):1105-11 [PMID: 19289445]
- Cufflinks/Cuffdiff  
Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010 May;28(5):511-5 [PMID: 20436464]
- Cuffdiff2 Differential analysis of gene regulation at transcript resolution with RNA-seq  
Nature Biotechnology 31, 46–53 (2013) [PMID: 23222703]
- TopHat-Fusion: an algorithm for discovery of novel fusion transcripts.  
Genome Biol. 2011 Aug 11;12(8):R72. [PMID: 21835007]
- STAR: ultrafast universal RNA-seq aligner.  
Bioinformatics\_ 2013 Jan 1;29(1):15-21. [PMID: 23104886]
- DESeq  
Differential expression analysis for sequence count data.  
Genome Biol. 2010;11(10):R106. [PMID: 20979621]  
DESeq2: [www.bioconductor.org/packages/2.13/bioc/html/DESeq2.html](http://www.bioconductor.org/packages/2.13/bioc/html/DESeq2.html) <http://www.ncbi.nlm.nih.gov/pubmed>

# ご参考サイト

## イルミナ

<http://www.illumina.com/landing/basespace-core-apps-for-rna-sequencing/>  
<http://res.illumina.com/documents/products/technotes/technote-basespace-rna-seq.pdf>  
[http://support.illumina.com/help/BaseSpace\\_App\\_RNAseq\\_help/RNAseq\\_Apps\\_Help.htm](http://support.illumina.com/help/BaseSpace_App_RNAseq_help/RNAseq_Apps_Help.htm)  
<http://www.illumina.com/applications/sequencing/rna.ilmn>

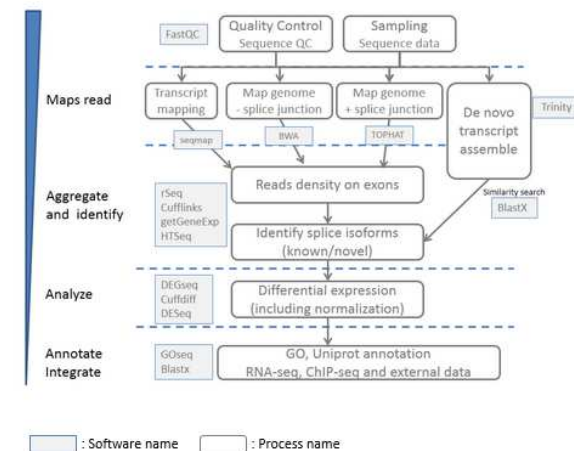
## 業界フォーラム(英語)

<http://seqanswers.com/>  
<https://www.biostars.org/>

## 日本語フォーラムサイト

<http://cell-innovation.nig.ac.jp/wiki/tiki-index.php>  
<http://qa.lifesciencedb.jp/>

ご参考：RNA-seq 典型プロセス と 典型ソフト



<http://cell-innovation.nig.ac.jp/wiki2>



Next is Now

Thank You

Questions?

