

2015年9月4日  
イルミナ サポートウェビナー

# 解析に適したリード前処理 を行うために

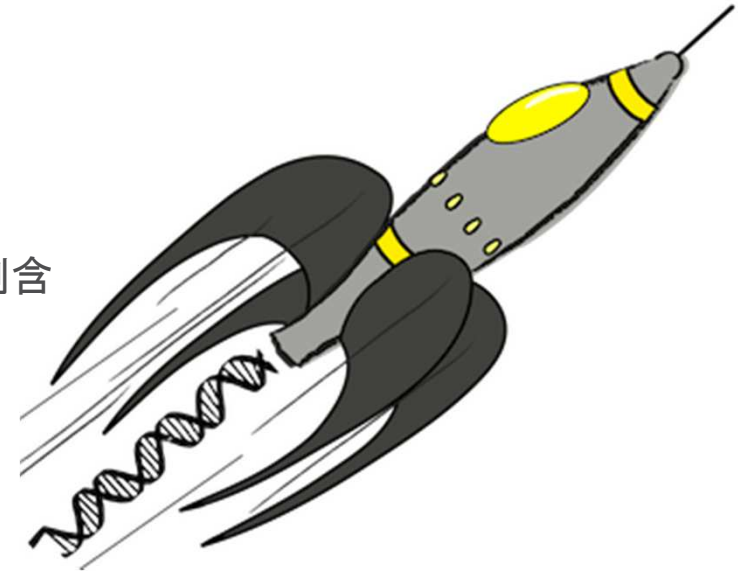


イルミナ株式会社  
バイオインフォマティクス  
サポートサイエンティスト  
癸生川絵里 (Eri Kibukawa)

※BaseSpaceアプリ: FASTQ toolkit /smallRNA/ FASTQC

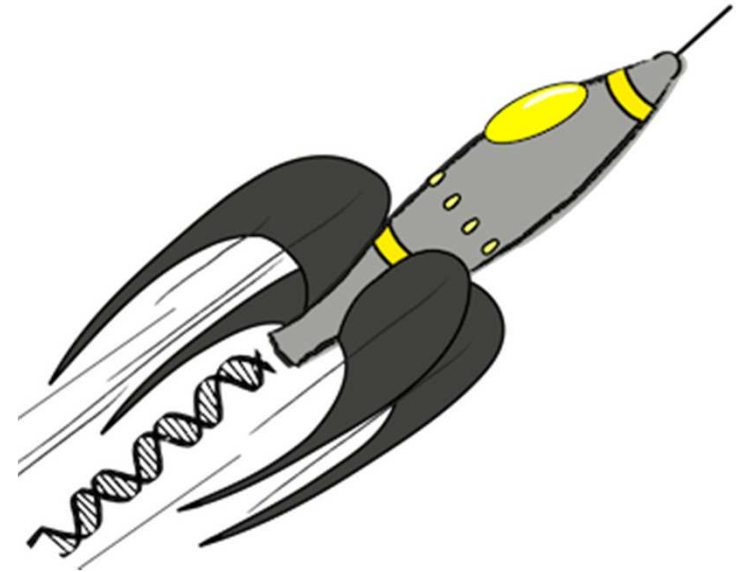
## 本日の内容

- イン트로ダクション
- アダプタートリミング  
※smallRNA 例含
- クオリティトリミング
- ダウンサンプリング
- リードの結合
- 手元のFASTQをトリミングするには

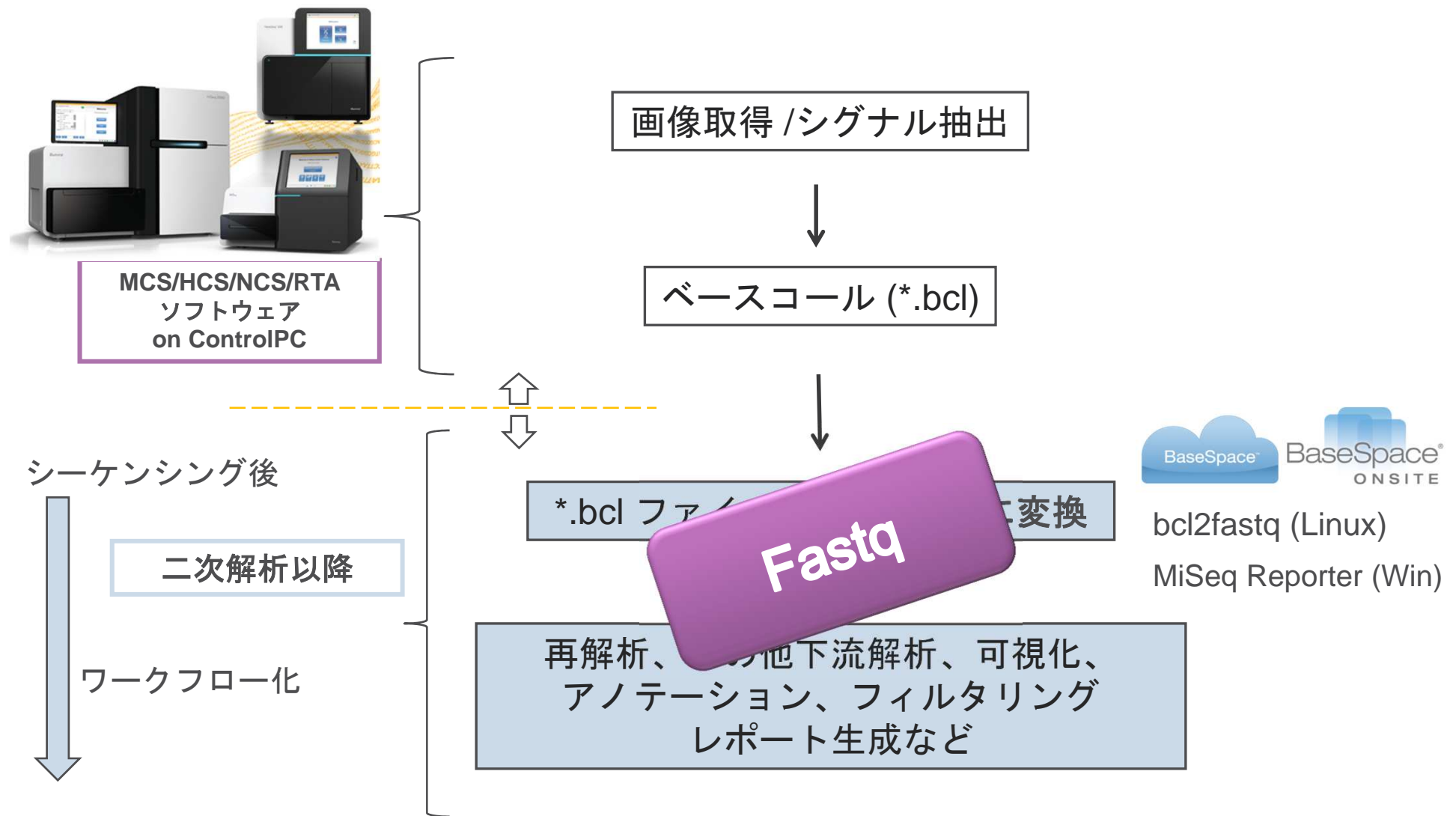


## 本日の内容

- イン트로ダクション
- アダプタートリミング
- クオリティトリミング
- ダウンサンプリング
- リードの結合
- 手元のFASTQをトリミングするには



# 装置からの解析フロー



# FASTQフォーマット

```
Header → @HWI-BRUNOP20X:994:B809UWABXX:1:1101:13501:2240 1:N:0:CTTGTA
Sequence → TGAAACCAGTGTTCTTAATTGGCATTTTACACACACACACACAGAATTTAAAAAAAAAATCAAAGG
+
Q-score → =55>7;?::BDADDD@EE88DCD?DFFEFFECBE6666BB=B;<;<-34:;<CB51>=BBEE>EE?

Header → @HWI-BRUNOP20X:994:B809UWABXX:1:1101:13660:2247 1:N:0:CTTGTA
Sequence → CCAAACATTAAGTAACTCTTAAAATGGCACACAGGTTTTAAAGCTATTGGTTTTTCCTTCCTAACT
+
Q-score → FFEDFBGEGGGGDFGEFFFFGGDF=FBFFFGGGE7CEEDEFBFBFGEEGF@FCDDFDFFEGFEAGF

Header → @HWI-BRUNOP20X:994:B809UWABXX:1:1101:13966:2183 1:N:0:CTTGTA
Sequence → TTGGGTAACTTGAATATAACATGGCTCCCTTGCTGTAAGCAAATGTTTTAGAGCTGAATTTTTCCT
+
Q-score → HHHHHEHHHHHHFHHHHHHHHHHHHHHHHHHHHHGGFHHHHHHHHHHFHHHFFEHFFEHFHHHHFHHHFF
```

# FASTQの生成場所・方法



MiSeqに内蔵されている。  
64bit Win に別途インストール  
も可能



BaseSpace®

BaseSpace®  
ONSITE

お使いのLinux server

**bcl2fastq2**



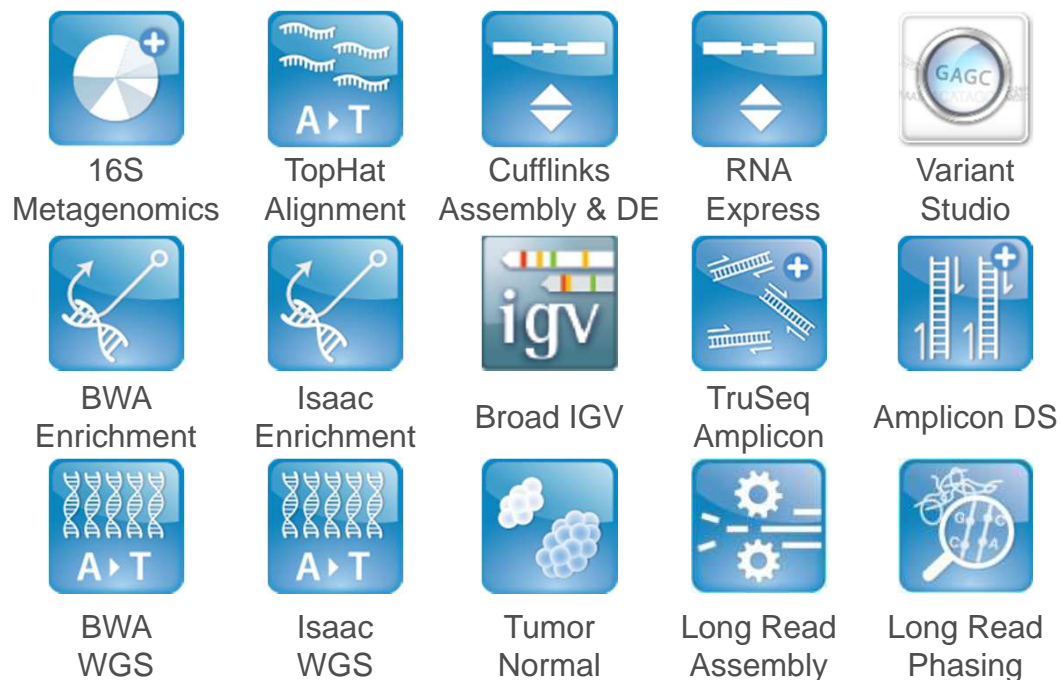
# アプリ (>60)



Categories	
Resequencing	Small RNA
Targeted Sequencing	De Novo Assembly
RNA-Seq	Gene Fusion Detection
ChIP-Seq	Methyl-Seq
Metagenomics	Tumor Normal
Variant Analysis	Differential Expression
Quality	Proteomics
Synthetic Long Reads	Visualization



## <イルミナコアアプリ>



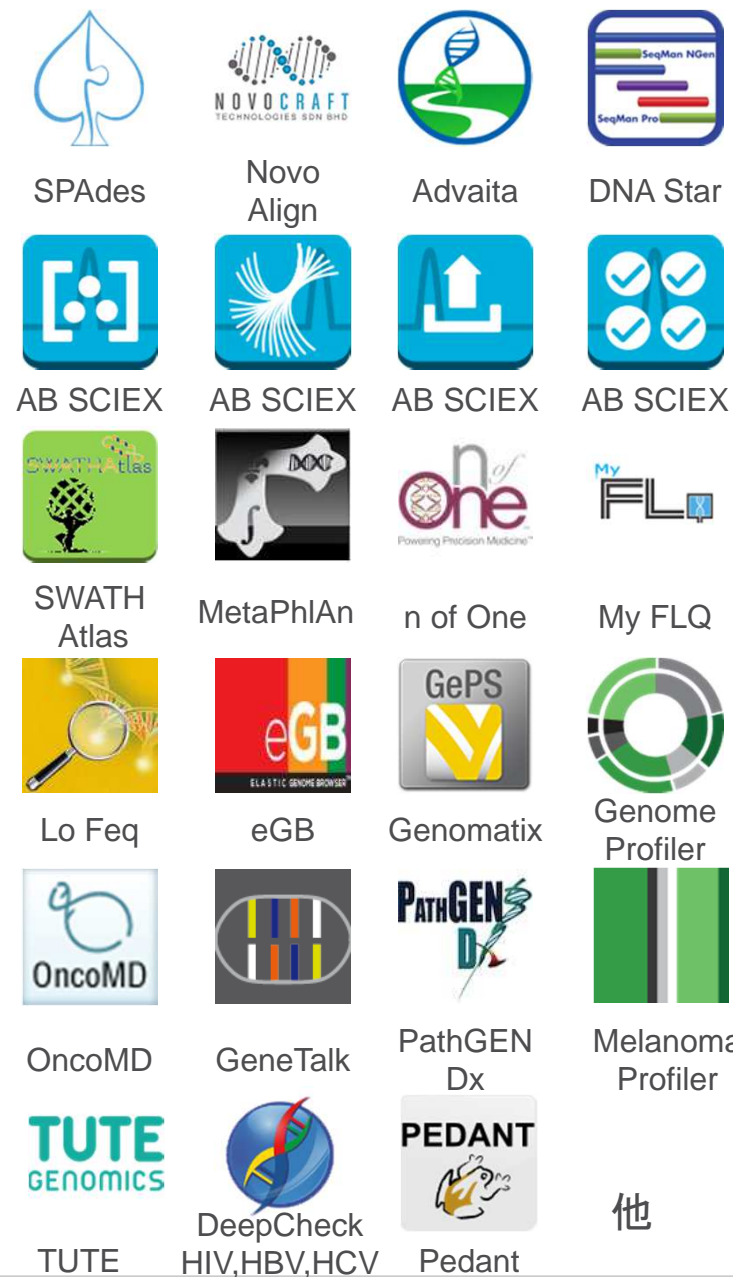
## <イルミナラボアプリ>



他

他

## <他社製アプリ>



他



# BaseSpace Labsアプリ (準サポート)

- ▶ 人気の機能をイルミナで素早くラップ/開発したツールをご提供.
- ▶ 一方、テストやドキュメント作成は低減
- ▶ テクニカルサポートの正式サポート対象ではなく、開発者へダイレクトにお問合せ戴けるご提供形態のアプリ([basespacelabs@illumina.com](mailto:basespacelabs@illumina.com)).

## FASTQ Toolkit



- ▶ Sub-sample reads
- ▶ Trim Adapters
- ▶ Trim Bases
- ▶ Ploy A/T trimming
- ▶ Quality Trimming
- ▶ Read Filtering
- ▶ Reverse Complement

## FastQC



- ▶ Perform QC of raw sequencing data.
- ▶ Determine adapter contamination

## VCAT v2.3



- ▶ Compare Variant Call Sets to standards
- ▶ Intersect variant call sets.

## SRA Import v0.0.3



- ▶ Import up to 25GB of sequencing data from SRA

## SRA Submission v0.0.3



- ▶ Deposit sequencing data in SRA.

他

# FASTQ Toolkit (FASTQツールキット)



## Adapter trimming (アダプタートリミング)

5'-また3'-それぞれ別にトリミングしたいアダプター配列を指定できる

## Base trimming (ベーストリミング)

5'-あるいは3'-端から、指定長分の塩基をトリミングすることができる

## Quality trimming (クオリティトリミング)

3'-端の低クオリティ配列をトリミングする用途向け. Qscore平均閾値を指定

## Poly-A/T trimming (Poly-A/T トリミング)

リード終端のPoly-A/T をトリム.

## Sub-sampling (サブサンプリング、またはダウンサンプリングとも呼称)

サンプルリードの一部を取り出し、より少ないサンプルリードセットをつくる

# FASTQ Toolkit (FASTQツールキット)



## Read filtering (リードフィルタリング)

最短/最長 塩基数や最大/最小 平均クオリティー値、最大/最小 GC含有率、低複雑度領域などの条件を指定し指定閾値外のリードを除外

## Modify reads (旧 Reverse complement)

相補鎖配列取得

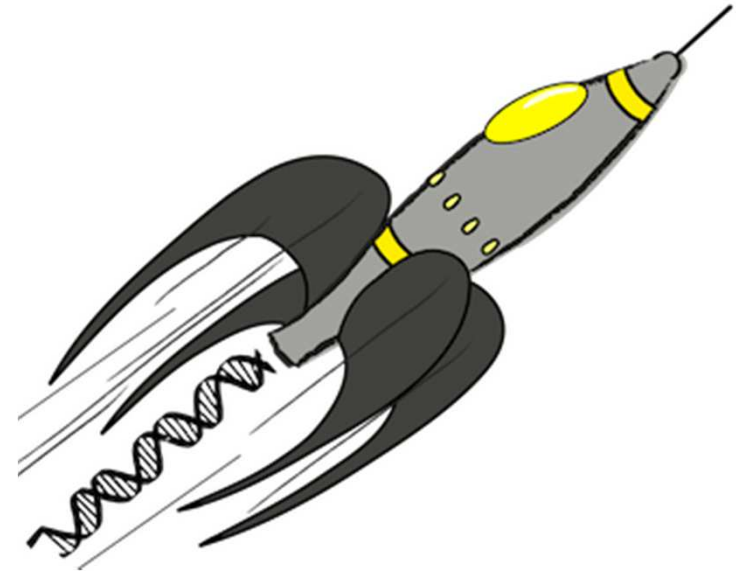
(Nexteraメイトペアリードからペアードエンドリード 方向への変換目的など)に加え、他ペアードエンドリードが1つのFASTQからR1, R2への振り分け

## Fix formats (フォーマット修正)

アップロードした FASTQヘッダやエンコード(Qscoreのオフセット値) 修正、ファイル名などが規約を満たしていない事によりBaseSpaceアプリが受け付けない場合に修正を試みるなど可能

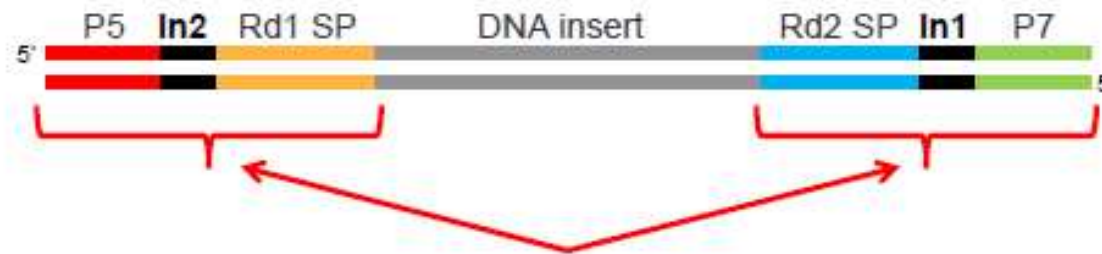
## 本日の内容

- イン트로ダクション
- アダプタートリミング
- クオリティトリミング
- ダウンサンプリング
- リードの結合
- 手元のFASTQをトリミングするには



# アダプターとは

## イルミナ ライブラリの構造



## イルミナシーケンサー専用オリゴヌクレオチドアダプター

- DNA インサート : 数百bpに断片化したDNA. 読みたい目的サンプル配列.
- P5, P7 : フローセルへの結合部位
- SP : シーケンシングプライマー結合部位
- In (Index) : 複数サンプル同時解析用のバーコード (目印配列)

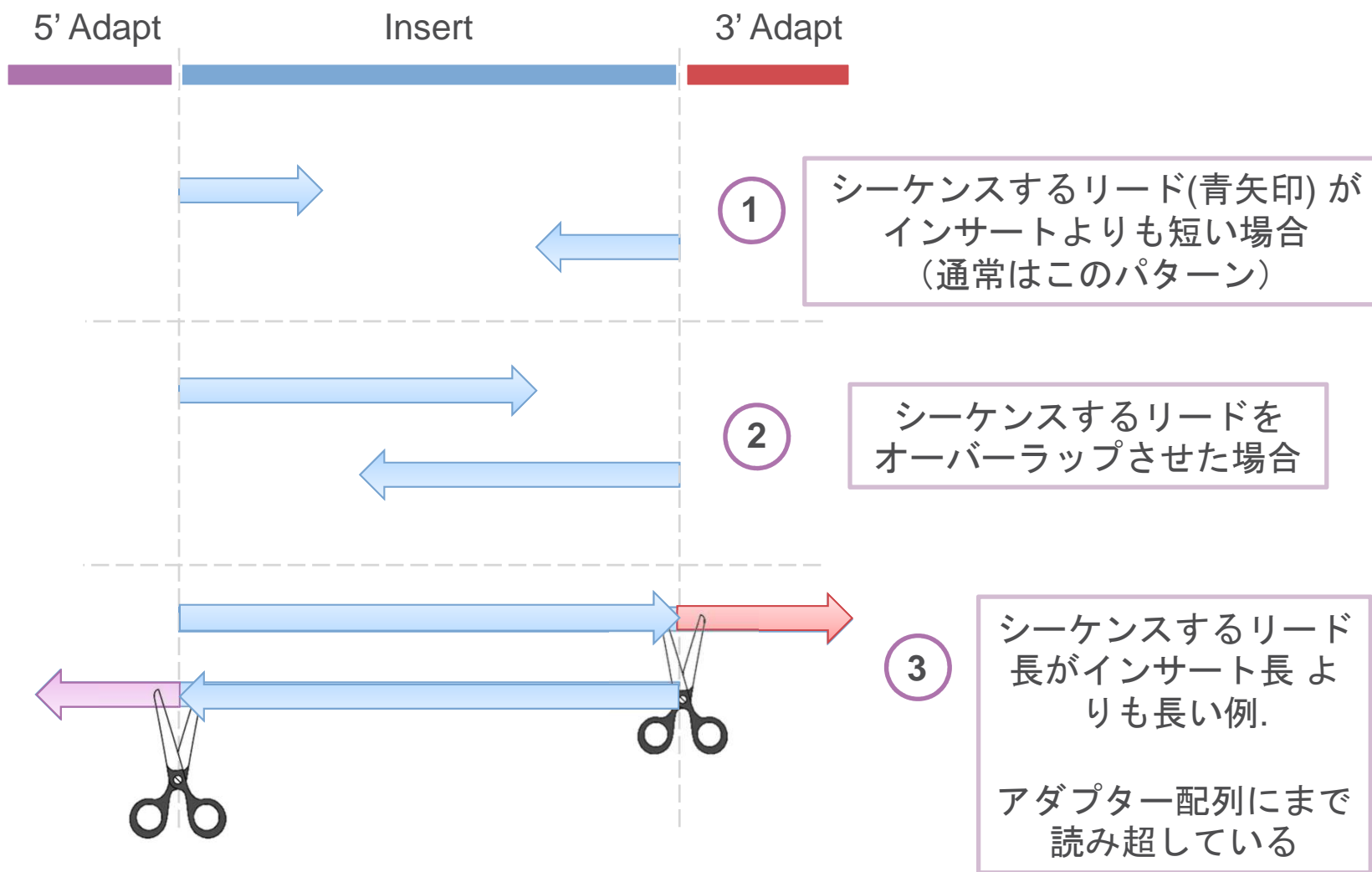
**ライブラリ = DNA インサート + 両端にそれぞれ別のアダプター**

イルミナシーケンサーでシーケンスするため、この構造をとるようにサンプル調整する

※ 詳しくは、弊社サポートウェビナー 2015/07/10 をご参考いただけます。  
**SBS (Sequencing By Synthesis) ケミストリーとは何か？**  
[http://www.illumina.co.jp/events/webinar\\_japan/support\\_webinar.ilmn](http://www.illumina.co.jp/events/webinar_japan/support_webinar.ilmn)

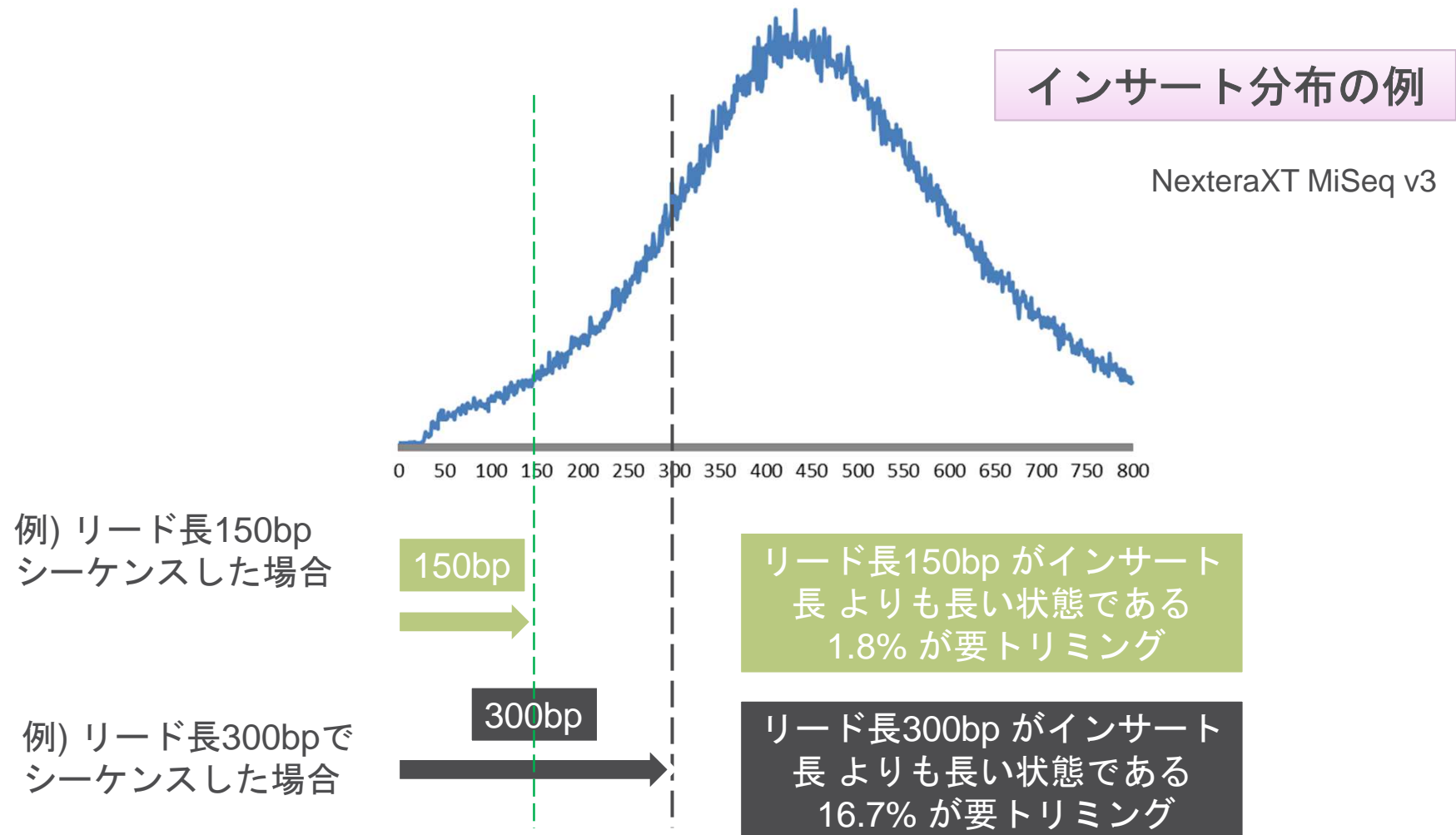
# インサート長とアダプタトリミング

アダプターとインサート配列からなるライブラリに対する、  
実際シーケンスしてリードとして得られる配列の位置関係のパターン





# インサート長の分布とアダプタートリミング



# アダプタートリミングの方法

## Adapter, AdapterRead2

[Header]  
IEMFileVersion  
Investigator Name  
Experiment Name  
Date  
Workflow  
Application  
Assay  
Description  
Chemistry

### トリミング

シーケンスから当該配列を除去（除去した分リード長が短くなる）

### [settings]

Adapter,.....

AdapterRead2,.....

Adapterのみに記載するとR1,R2ともにその配列でトリミングがされます（Nextera）

### [Reads]

151

151

### [Settings]

Adapter	AGATCGGAAGAGCACACGTCTGAACTCCAGTCA
AdapterRead2	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

### [Data]

Sample_ID	Sample_N	Sample_P	Sample_V	Sample_P	Description
test	test		test		

# アダプタートリミングの例

アダプター配列  
マッチ > 90%  
(デフォルト)

## ビフォー

```
@M00000:71:000000000-D00LW:1:1101:16265:1658 1:N:0:1
ACTCTGCGTTGCGCTTCTGCTCGGCCTCCAGCTCACCTCCCTGTCTCTTATACACATCTCCGAGCCCA
+
BCCCCFFCCBCCGGGGGGGGGGGGGGGGGGGGHHHHHHHHHHHHHHHHGHHHHHHHHHHHHHHHHHHHHGGGGGGH
```

## アフター

```
@M00000:71:000000000-D00LW:1:1101:16265:1658 1:N:0:1
ACTCTGCGTTGCGCTTCTGCTCGGCCTCCAGCTCACCTCC
+
BCCCCFFCCBCCGGGGGGGGGGGGGGGGGGGGHHHHHHHHHHHHHHHHG
```



当該アダプター配列の初頭から以降がトリムされる

# アダプターマスキング

## MaskAdapter, MaskAdapterRead2

[Header]	
IEMFileVersion	4
Investigator Name	Mr.X
Experiment Name	Example
Date	#####
Workflow	GenerateFastq
Application	GenerateFastq
Assay	TruSeq LT
Description	Example
Chemistry	Default
[Reads]	
	151
	151
[Settings]	
Adapter	AGATCGG
AdapterRead2	AGATCGG
[Data]	
Sample_ID	Sample_1
test	test

除去するのではなく、配列をNでマスクして残すこともできる。

(マスクしたNのqscoreは一律に“#”で差し替えられる)

[settings]のオプション名を以下で記載 or 書き換え

MaskAdapter,.....

MaskAdapterRead2,.....

※MiSeq Reporter、BaseSpace、bcl2fastq2等 利用時のサンプルシート設定

# アダプターマスキングで実行した例

## ビフォー

```
@M00000:71:000000000-D00LW:1:1101:16265:1658 1:N:0:1
ACTCTGCGTTGCGCTTCTGCTCGGCCTCCAGCTCACCCTCCCTGTCTCTTATACACATCTCCGAGCCCA
+
BCCCCFFCCBCCGGGGGGGGGGGGGGGGGGGHHHHHHHHHHHHHHGHHHHHHHHHHHHHHHHHHHHHHGGGGGGH
```



## アフター

```
@M00000:72:000000000-D00LW:1:1101:16265:1658 1:N:0:1
ACTCTGCGTTGCGCTTCTGCTCGGCCTCCAGCTCACCCTCCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+
BCCCCFFCCBCCGGGGGGGGGGGGGGGGGGGHHHHHHHHHHHHHHG#####
```


アダプター配列を含むアダプター配列以降の塩基をNでマスクし、  
クオリティスコアは一律 2 (#)で置換

# BaseSpaceでトリミング目的に使えるツール

## FASTQ Toolkit

BaseSpace

illumina



FASTQ Toolkit  
BaseSpace Labs

Analysis Name:

FASTQ Toolkit

Input Sample(s) to Process:

Select Sample(s)

Save Processed Sample(s) to:

Select Project(s)

Add this string to the output sample name(s):

▶ Sub-sampling

▶ Adapter Trimming

▶ Base Trimming

▶ Poly-A/T Trimming

▶ Quality Trimming

▶ Read Filtering

▶ Modify Reads

BaseSpace Labs:

☐ I acknowledge that I am using it AS IS. Illumina has no obligation to provide support, training, or other resources. (iv) Illumina has no liability for any damages, including consequential damages, arising from the use of the software.

▼ Adapter Trimming

Use these parameters if you want to trim adapter sequences. The adapter sequence can be specified separately for the 5'- and 3'-end and is a required input for adapter trimming. In addition to the adapter sequence, you can specify a stringency value that determines how similar a sequence has to be in order to be identified as an adapter sequence.

Adapter trim stringency (0.01-0.99):  
0.9

Select an adapter to trim or specify the adapter sequence(s) below:  
None selected

Adapter sequence(s) to trim from the 5'-end:

Adapter sequence(s) to trim from the 3'-end:  
CTGTCTCTTATACACATCTCCGAG

Trim Ns from the 3'-end before identifying adapters:



## その他アダプタートリミングに使える3rd-partyツールの一例

ツール名	配布場所
Trimmomatic	<a href="http://www.usadellab.org/cms/?page=trimmomatic">http://www.usadellab.org/cms/?page=trimmomatic</a>
FASTX toolkit	<a href="http://hannonlab.cshl.edu/fastx_toolkit/">http://hannonlab.cshl.edu/fastx_toolkit/</a> (FastQ clipper)
Seq-Prep	<a href="https://github.com/jstjohn/SeqPrep">https://github.com/jstjohn/SeqPrep</a>
Cut-Adapt	<a href="https://code.google.com/p/cutadapt/">https://code.google.com/p/cutadapt/</a>
PEAT	<a href="https://github.com/jhhung/PEAT">https://github.com/jhhung/PEAT</a> アダプター配列そのものを指定せずにトリミングができる (PEの重なりから判別するため、PE必須)

参考 : <http://omictools.com/adapter-trimming-c402-p1.html>

# なぜアダプター配列トリムを検討するのか？



BWA  
Enrichment  
V2.1

1

アライメントできるリード量が増える  
場合がある

BWA  
(backtrace)

Sample	Sample Name	Total Aligned Reads	Percent Aligned Reads
1	NA12892	354,882	77.4%
2	NA12892-trim	450,007	98.2%

ただし: 使用しているアライナープログラムによる

BWA  
(mem)

Sample	Sample Name	Total Aligned Reads	Percent Aligned Reads
1	NA12892	457,611	99.8%
2	NA12892-trim	457,542	99.8%

# なぜアダプター配列トリムを検討するのか？



Velvet de novo Assembly  
BASESPACE LABS

2

## 例えばアセンブル結果の向上

Assembly metrics	Before adapter trimming	After adapter trimming
N50	21	29,791
Maximum contig	553	174,326
Assembly length	18,497,207	4,876,437
Number of contigs	1,387,508	1,115

2 x 250bp, E.coli (Nextera XT)



# なぜアダプター配列トリムを検討するのか？



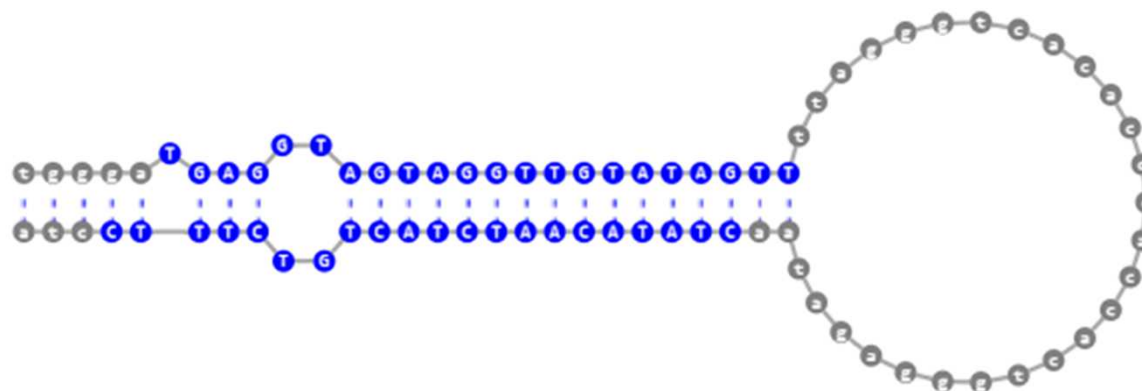
Small RNA v1.0

## 3

## Small RNA のワークフローで必要となる

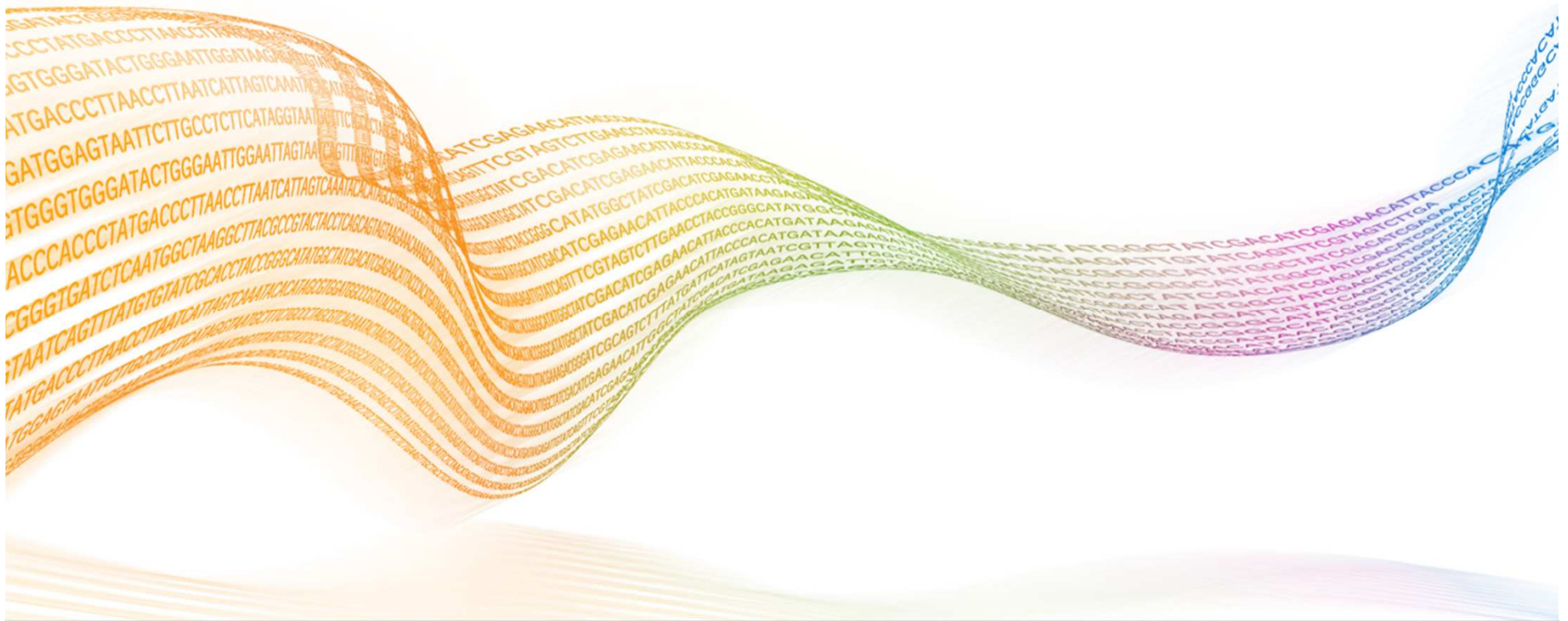
smallRNA解析では通常非常に短い配列を対象とするため、シーケンシングのリード長の方が、smallRNAのインサート長よりも、短くなる。したがって、アダプタートリミングが定常処理として必要となってくる。

(例 ヒト miRNAだと例えば分布ピークが 22bpなど)



アダプタートリミングが必用となる例：

## Small RNA 解析



# Small RNA のワークフロー

## MiSeqの場合



- 1 Illumina Experiment ManagerウィジェットでSampleSheetを作成する際、“smallRNA”ワークフローを選択する。  
シーケンシングを開始する。
- 2 生成されたFASTQファイルは自動でアダプタトリム済みとなる。  
明示的にサンプルシートには記載なくとも**デフォルトでトリムが適用されている**。  
*TruSeq small RNA adapter (TGGAATTCTCGGGTGCCAAGG)*  
他のキットを使用している場合は明示的にサンプルシートに記載すれば適用される。
- 3 MiSeq ReporterではsmallRNAのワークフローによりレポート生成まで自動実行される。  
途中で出力されたFASTQは、アダプタトリム済みのため、  
BaseSpaceにアップロードするなどしてさらに後続の解析にそのまま使う事が可能。



# BaseSpace Small RNA v1.0 アプリ

※ アダプタートリム済みのリードが必用



Small RNA v1.0

## 対応装置データ

- ▶ HiSeq 2500/3000/4000
- ▶ NextSeq 500
- ▶ MiSeq

## 対応ライブラリ調整キット

- ▶ TruSeq Small RNA

## 対応ゲノム

- ▶ Human HG19
- ▶ Mus musculus
- ▶ Rattus norvegicus

## 機能

- ▶ Alignment
- ▶ Classification of miRNAs, isomiRs, and piRNAs
- ▶ Novel miRNA discovery
- ▶ miRNA Precursor discovery
- ▶ Differential Expression of miRNAs, precursor groups, miRNA families, and piRNAs

## 内包ソフトウェアバージョン

- ▶ Isis (Analysis Software)—2.5.52.11
- ▶ Samtools 0.1.19-isis-1.0.2
- ▶ Bowtie (Aligner) 0.12.8
- ▶ miRDeep\* 3.2
- ▶ DESeq2 1.0.17

# Small RNAのワークフロー (GenerateFastq) HiSeq/ NextSeq の場合



- ① smallRNAは装置からBaseSpace直アップロードの際は、留意が必要※  
アダプター配列を自動トリムされないようにする必要がある  
サンプルートはGenerateFASTQを指定、かつアダプタを記入しないなど(HiSeq)
- ② FASTQ Toolkit アプリなどでアダプタートリムを行っておく
- ③ トリム済みのFASTQをsmallRNA v1.0アプリの入力に供する

※ BaseSpaceにおいてGenerateFastq でアダプタートリムの指定を行うと32 bp よりも短い配列は一律に Nでマスクされるため。

# Small RNAのリードを Fastq toolkitでトリムする



- 1 ProjectエリアのLaunch appボタンなどから“FASTQ Toolkit”アプリを起動



- 2 Select Samples で入力サンプル(= fastq)を選択し “Add a string to the output sample name(s)”にファイル名に別名を付けるための文字列を入力

Analysis Name:

FASTQ Toolkit v1.0 04/29/2015 4:40:39

Input Sample(s) to Process:

Select Sample(s):

subHuBr1

Save Processed Sample(s) to:

Select Project(s):

smallRNA-test

Add this string to the output sample name(s):

trim

例: 上記のようにtrimを入れておくと、トリム後のサンプル名(fastqファイル名)が“subHuBr1trim”となる。  
オリジナルとの区別のため。

# TruSeq Small RNAのリードを Fastq toolkitでトリムする

3

トリムしたいアダプター配列を選ぶ:

“Adapter trimming” > “Adapter sequence(s) to trim from the 3’ end”:

“TGGAATTCTCGGGTGCCAAGG” (This is the TruSeq smallRNA adapter)

## ▼ Adapter Trimming

Use these parameters if you want to trim adapter sequences. The adapter sequence can be specified separately for the 5'- and 3'-end and is a required input for adapter trimming. In addition to the adapter sequence, you can specify a stringency value that determines how similar a sequence has to be in order to be identified as an adapter sequence.

Adapter trim stringency (0.01-0.99):

0.9

ドロップダウンから選べるキットもある

Select an adapter to trim or specify the adapter sequence(s) below:

None selected

Adapter sequence(s) to trim from the 5'-end:

Adapter sequence(s) to trim from the 3'-end:

TGGAATTCTCGGGTGCCAAGG

Trim from the 5'-end before identifying adapters:

☐

# TruSeq Small RNAのリードを Fastq toolkitでトリムする

4

最低リード長を入力

“Read Filter” > “Minimum Read length: 15” (変更可能)

▶ Quality Trimming

▼ Read Filtering

Use these parameters if you want to filter reads. Paired-end reads are only filtered (removed from the sample) if both reads are filtered out. Otherwise, the filtered mate is replaced by a sequence of Ns (number of Ns will be the minimum read length) to keep the order of pairs in the FASTQ files (necessary for many secondary analysis tools).

Minimum read length:

15

Maximum read length:

Minimum mean quality score:

Maximum mean quality score:

*Note, that leaving as default will result in conversion of sequences <32bp to “N” strings*

# TruSeq Small RNAのリードを Fastq toolkitでトリムする

## 5 “BaseSpace Labs Apps” Agreement にチェックを入れて承諾する

Maximum GC content:

Minimum sequence complexity:

Only keep reads passing filters:



► Modify Reads

BaseSpace Labs:

- ☒ I acknowledge and agree that (i) this is a BaseSpace Labs App, (ii) I am using it AS-IS without any warranty of any kind, (iii) Illumina has no obligation to provide any technical support for this App, and (iv) Illumina has no liability for my use of this App, including without limitation, any loss of data, incorrect results, or any costs, liabilities, or damages that result from use of this App.

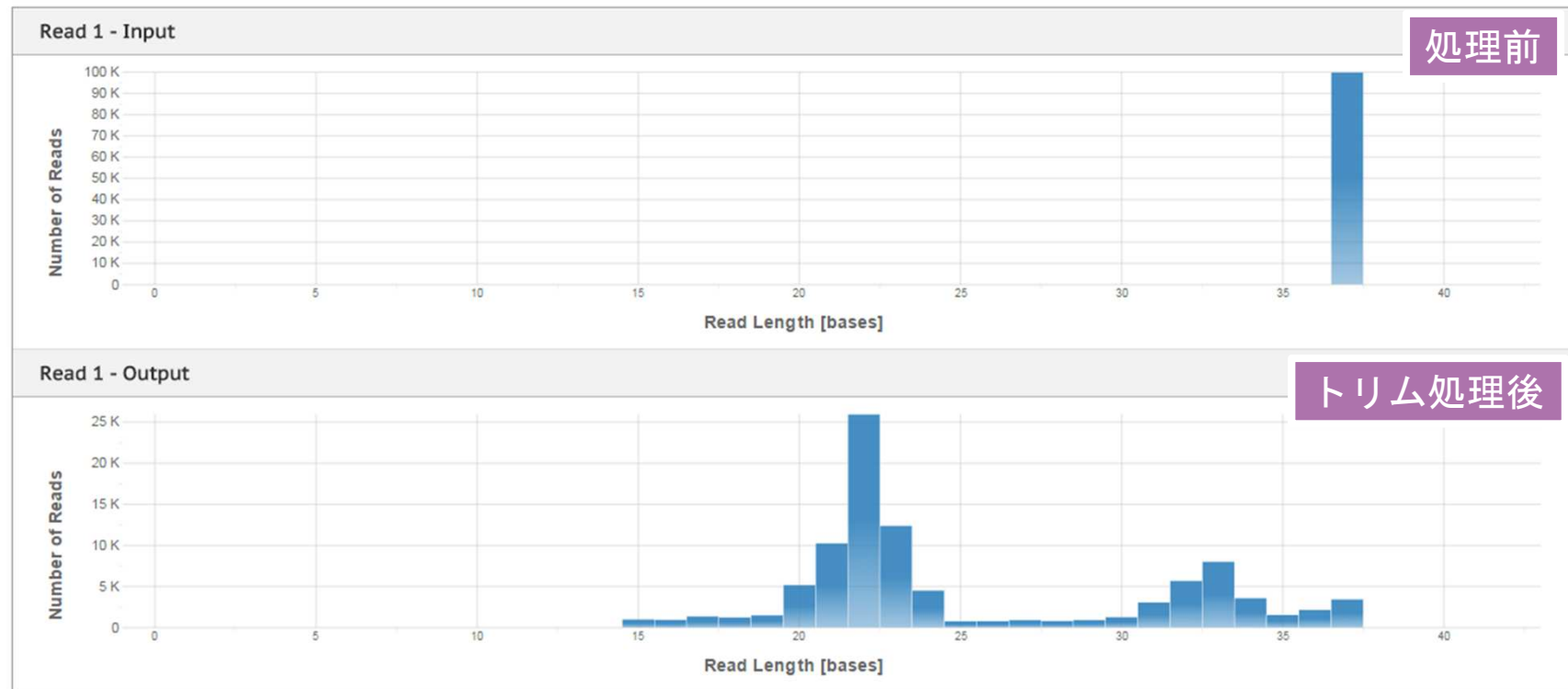
AS-ISでご使用いただくことの明示的ご了承

**Continue**ボタンを押し、実行を開始する



# TruSeq smallRNA のリードを Fastq toolkitでトリム 結果のレポート (ビフォーアフター)


Read Length Distributions



Note: Reads that only contain Ns (no-calls) in the output sample are represented with length zero in the charts above.


(レポートの一部抜粋)

# BaseSpace Small RNAアプリ

**Small RNA v1.0**  
ILLUMINA  
[Launch](#)

Free

Precursor: **hsa-let-7a-1**



[Save as PNG](#)

**hsa-let-7a-1 in miRBase**  
5' Counts: 203771  
3' Counts: 129  
5'/3' ratio: 1579.62

```
tggaTGAGGTAGTAGTTGTATAGTTtaggggtcacaccaccactggagataaCTATACAATCTACTGCTTTCcta
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
...TCTGAGGTAGTAGTTGTATAGTT..... 6
....TGAGGTAGTAGTTGTATAGTT..... 196698
....TGAGGTAGTAGTTGTATAGT..... 7061
....TGAGGTAGTAGTTGTGTTT..... 6
.....CTATACAATCTACTGCTTTC... 33
```

[Overview](#)

### Description

The BaseSpace Small RNA v1.0 app analyzes small RNA samples. The Small RNA app supports TruSeq Small RNA Sample Preparation Kits and assumes that reads have been properly adapter-trimmed. The app aligns reads against four reference databases (abundant, mature miRNA, other RNA, and genomic) and outputs hits to mature miRNAs, isomiRs, and piRNAs. Optionally, the app performs novel precursor discovery and pairwise differential expression analysis. Pairwise differential expression analysis identifies differentially expressed miRNAs, precursor groups, miRNA families, and piRNAs for each pair of sample groups.

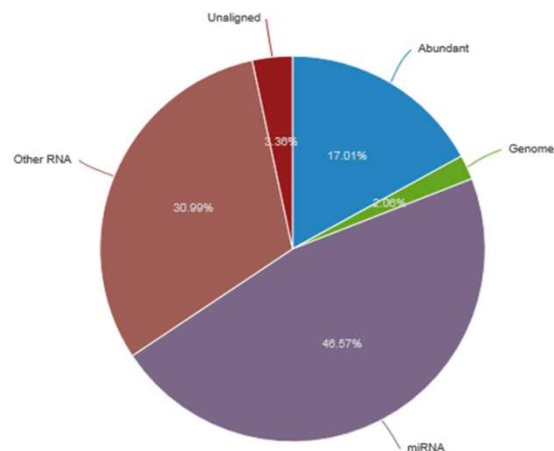
This application is for research use only.

[Support](#)

**Categories**  
[Native](#)  
[Small RNA](#)  
[RNA-Seq](#)  
[Differential Expression](#)

**Small RNA v1.0**  
ILLUMINA  
[Launch](#)

# Small RNA アプリ結果のレポート



hsa-miR-20a-3p
hsa-miR-22-3p
hsa-let-7a-5p
hsa-miR-9-5p
hsa-miR-27b-3p
hsa-let-7f-5p
hsa-miR-143-3p
hsa-miR-127-3p
hsa-miR-126-5p

The full list is available at Hits.txt.

## IsomiRs (Known Precursor)

A read is counted as isomiR if it is a subsequence on the same strand of a pre-miRNA sequence. The precursor sequences were obtained from miRBase. An isomiR hit is labeled as "start position}\_{sequence read}\_{precursor ID}", where the start position is 0-based. Mismatches are allowed. IsomiR hits are filtered to remove artifacts of PCR amplification.

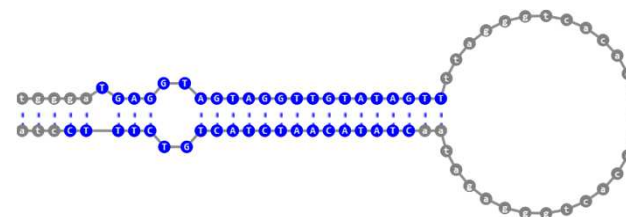
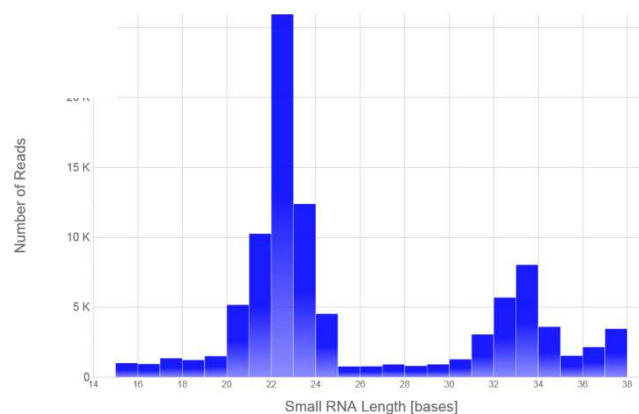
Number of IsomiRs (Known Precursor) with Reads	Total Number of Reads
464	22,967

## ces

Reads are listed for each type. If there are less than 10 sequences with a certain length, they are not listed.

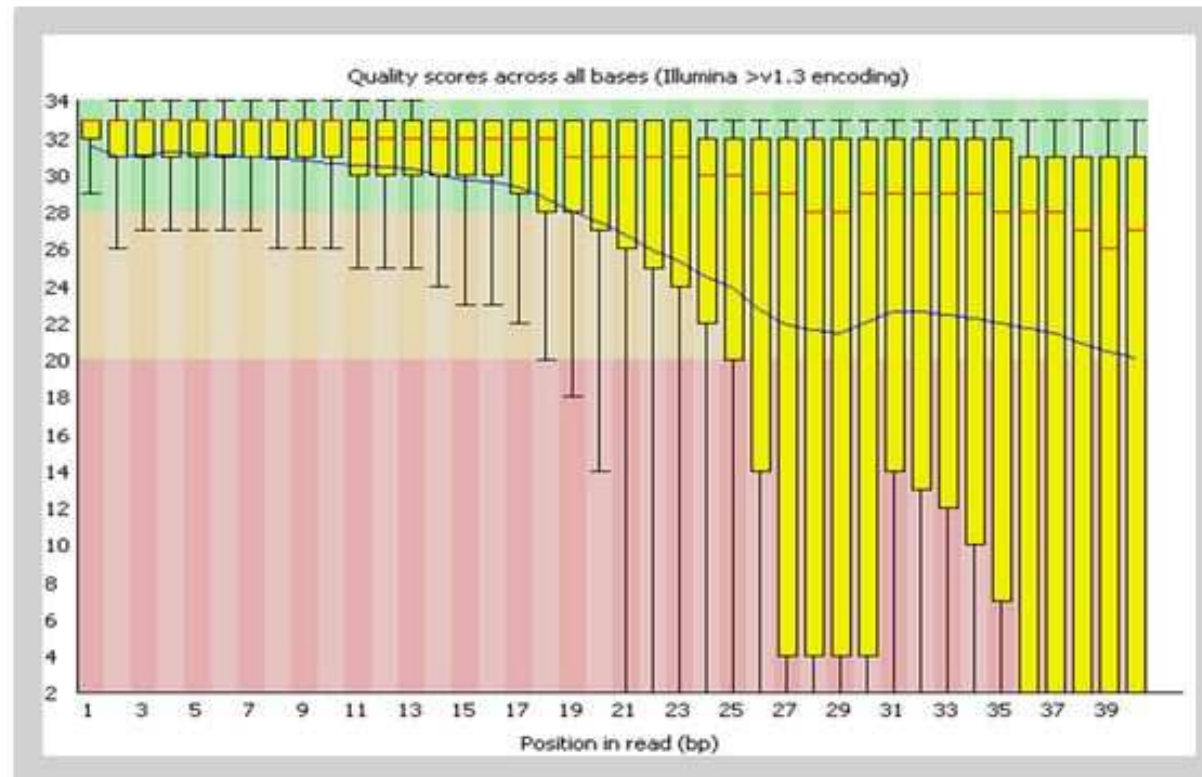
Records are counted. A read must align to the start of a reference sequence and have the same length of the reference sequence. No mismatches are allowed.

Reads	Total Number of Reads
281	22,967



# Small RNA

このFASTQリードはトリムされたものか？  
– FastQCアプリ



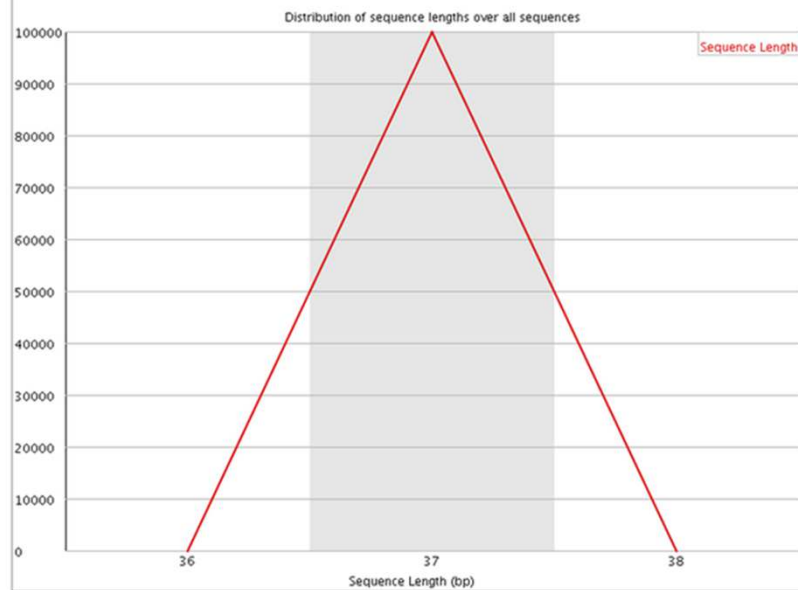
# Small RNA

このFASTQリードはトリムされたものか？  
– FastQCアプリ



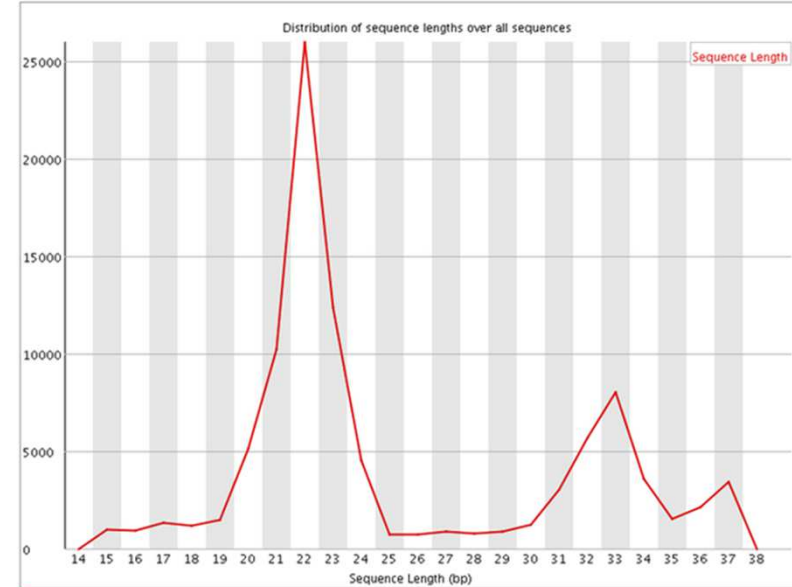
トリムされていない

## Sequence Length Distribution



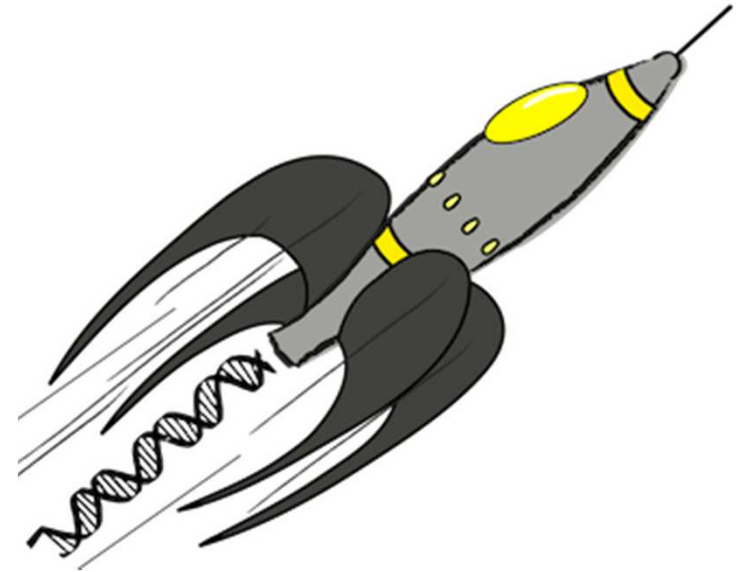
トリムされている

## Sequence Length Distribution



## 本日の内容

- イン트로ダクション
- アダプタートリミング
- クオリティトリミング
- ダウンサンプリング
- リードの結合
- 手元のFASTQをトリミングするには



# クオリティースコア(qscore) によるトリミング



- とはなにか？  
3'末端のクオリティの平均に基づきトリミングする
- 3'末端からのスライディングウィンドウのアプローチをとり、枠をスライドさせながら平均クオリティが閾値を下回ったときに以降をトリムするものが多い
- どのような時に行うものなのか？  
後続の解析でベースコールのクオリティがシビアに影響するような解析の場合。  
例えば- de novoアセンブリ、リードの結合、リードからの分類(メタゲノム解析など)
- 逆に、どのようなときは使われないもの？
  - リシーケンシング解析. ほとんどのアライメントツールは塩基のqscoreも計算に入れており (i.e. BWA, Isaac)、末端に低 qscore 配列がある場合はソフトウェア的に省く処理が実装されている等

# Qスコアによるトリミング

## GenerateFastq in MSR/ BaseSpace /bcl2fastq2)

[Header]			
IEMFileVersion	4		
Investigator Name	Mr.X		
Experiment Name	Example		
Date	4/05/2015		
Workflow	Resequencing		
Application	Resequencing		
Assay	TruSeq LT		
Description	Example		
Chemistry	Default		
[Reads]			
	151		
	151		
[Settings]			
Adapter	AGATCGGAAGAGCACACGTCTGAA		
Adapter_Read2	AGATCGGAAGAGCGGTCTGTAGGGA		
QualityScoreTrim	20		
[Data]			
Sample_ID	Sample_Name	Sample_Plate	
test	test		test

QualityScoreTrim

[settings]

QualityScoreTrim,<qualityScore>



# Qスコアによるトリミングの例

## QualityScoreTrim,20

ビフォー

```
@M00000:72:000000000-D00LW:1:1101:22420:18334 1:N:0:1
CACCAAGGGCCTGGGGTGTCAATGGCGGGGCTTGTGACTGCACAAAAGGGCCTCCCGCAGGGGCTCCCGCC
+
BBBBBBFBBBBBGGGGEEFGGGHHHHGGG00>10B355@BB3@3BG1?E1///1B11//////////?/////
```



アフター

```
@M00000:72:000000000-D00LW:1:1101:22420:18334 1:N:0:1
CACCAAGGGCCTGGGGTGTCAATGGCGGGGCTTGTGACTGCACAAAAGG
+
BBBBBBFBBBBBGGGGEEFGGGHHHHGGG00>10B355@BB3@3BG1?E
```




Q	ASC
13	.
14	/
15	0
16	1
18	3
20	5
22	7
25	9
30	?
31	@
32	A
33	B

# BaseSpace アプリによる Quality トリミング FASTQ Toolkit

BaseSpace

illumina



FASTQ Toolkit  
BaseSpace Labs

Analysis Name:

FASTQ Toolkit v1.0 03/17/2015 7:07:19

Input Sample(s) to Process:

Select Sample(s):

Save Processed Sample(s) to:

Select Project(s):

Add this string to the output sample name(s):

▶ Sub-sampling

▶ Adapter Trimming

▶ Base Trimming

▶ Poly-A/T Trimming

▶ Quality Trimming

▶ Read Filtering

▶ Modify Reads

Trim the 3'-end of reads with quality level:

BaseSpace Labs:

☐ I acknowledge and agree that (i) this is a BaseSpace Labs App, (ii) I am using it AS-IS without any warranty of any kind, (iii) Illumina has no obligation to provide any technical support for this App, and (iv) Illumina has no liability for my use of this App, including without limitation, any loss of data, incorrect results, or any costs, liabilities, or damages that result from use of this App.

This app is free.

→

contact us

## Quality トリミング

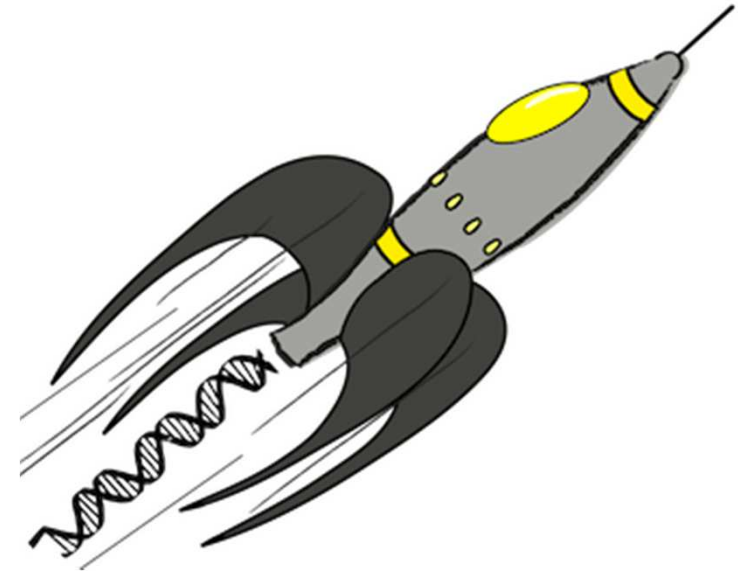
### 3<sup>rd</sup>- party ツール例

ツール名	URL
Trimmomatic	<a href="http://www.usadellab.org/cms/?page=trimmomatic">http://www.usadellab.org/cms/?page=trimmomatic</a>
Trim-Galore	<a href="http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/">http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/</a>
FASTX toolkit	<a href="http://hannonlab.cshl.edu/fastx_toolkit/">http://hannonlab.cshl.edu/fastx_toolkit/</a> (FastQ clipper)

参考 : <http://omictools.com/adapter-trimming-c402-p1.html>

## 本日の内容

- イン트로ダクション
- アダプタートリミング
- クオリティトリミング
- **ダウンサンプリング**
- リードの結合
- 手元のFASTQをトリミングするには



# ダウンサンプリング (サブサンプリング)




- とはなにか？
  - リード量が多すぎるときに一部のリードを取り出す（サブセットをつくる）
- なぜあえてサンプリングによりリード量を減らすのか？
  - トラブルシュートなどで素早くリードを検分(QC)したいとき、全リードで分析するとあまりに大量で解析時間がかかるため、負荷軽減、時間短縮をねらって.
  - 解析環境や解析ツール、サンプル特異性によって解析系が大量リードの処理に耐えない場合がある. このエラーを回避し解析を進めるために入力リード量を減らす必要が生じる場合がある.
    - 例) メモリー不足で落ちる、ディスク領域が足りないなど
  - BaseSpaceのアプリでも入力データ量の制限を明記しているものがある. こういったアプリや3rd-partyツールの入力制限に合わせるため.
  - 入力量で解析結果がどのように影響されるかなどの解析条件検討.
- イルミナでサブサンプリングをするには
  - BaseSpace FASTQ toolkit アプリ

# BaseSpace App: FASTQ Toolkitによるサブサンプリング

← BaseSpace

illumina

 FASTQ Toolkit  
BaseSpace Labs

Analysis Name:

FASTQ Toolkit v

Input Sample(s) to Process:

Select Sample(s)

Save Processed Sample(s) to:

Select Project(s)

Add this string to the output sample name(s):

▶ Sub-sampling

▶ Adapter Trimming

▶ Base Trimming

▶ Poly-A/T Trimming

▶ Quality Trimming

▶ Read Filtering

▶ Modify Reads

BaseSpace Labs:

☐ I acknowledge am using it AS-is. Illumina has no obligation to provide support without limitation of liabilities, or damages.

▼ Sub-sampling

Sub-sampling is required when only a subset of the sample can be processed by an application (e.g. de novo assembly with memory constraints) or it is not necessary to process a full sample (e.g. for validating an approach at varying levels of genomic coverage). Samples can be sub-sampled to a user defined number of reads or number of bases, or to a specified fraction of the input sample (e.g. 50% of input sample) - again, based on reads or bases.

Maximum number of FASTQ entries to keep:

Maximum percentage of FASTQ entries to keep:

Maximum number of bases to keep:

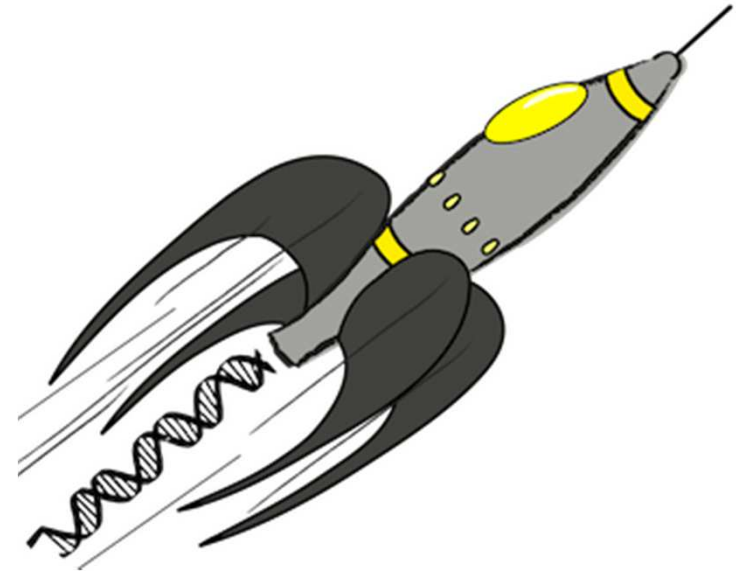
Maximum percentage of bases to keep:

Choose reads at random when sub-sampling instead of taking the first n reads:

☒

## 本日の内容

- イン트로ダクション
- アダプタートリミング
- クオリティトリミング
- ダウンサンプリング
- リードの結合
- 手元のFASTQをトリミングするには



# リードのマージ (結合、join、stitch など呼称さまざま)



## ○ とはなにか？

- 重複領域を頼りにリードをつなぎ合わせること

狭義では、ペアードエンドのR1とR2をつなぎ合わせることに

通常はある程度クオリティーの良い塩基のオーバーラップが一定長以上あることを条件とし、つなぎあわせる処理を行う (Q15以上の塩基が連続25bp以上など)

## ○ どういう時に行うものなのか？

- リードを長くすることが大切な場合
- indel 検出の向上に使えることもある
- 以降の解析ツールがシングルエンドしか受け付けない様なものの場合 (一部のメタゲノム解析ツールなど)
- ほとんどのリードがオーバーラップするようなデザインで読んだもの

## ○ 逆に、適さないときは？

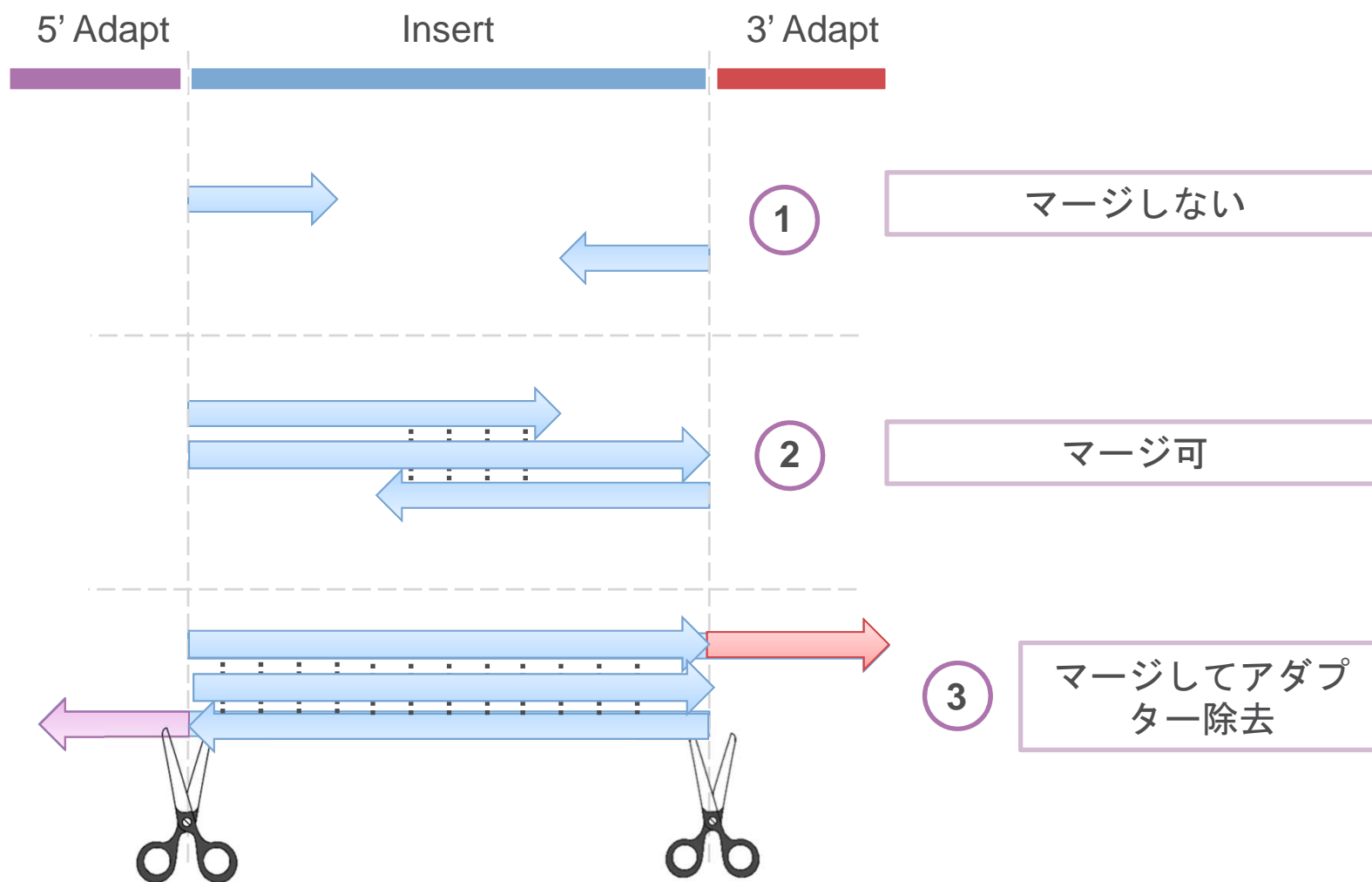
- クオリティーの良い塩基のオーバーラップがない
- 一部のリードしかオーバーラップがない場合 (設計外)
- オーバーラップ領域にリピート配列が予想されるとき

## ○ イルミナでリードのマージをするには

- MiSeq ReporterではStitch Readという機能でR1,R2のマージ可能 (一部ワークフロー)



# リードマージの概念図



## リードマージができるツールの一例

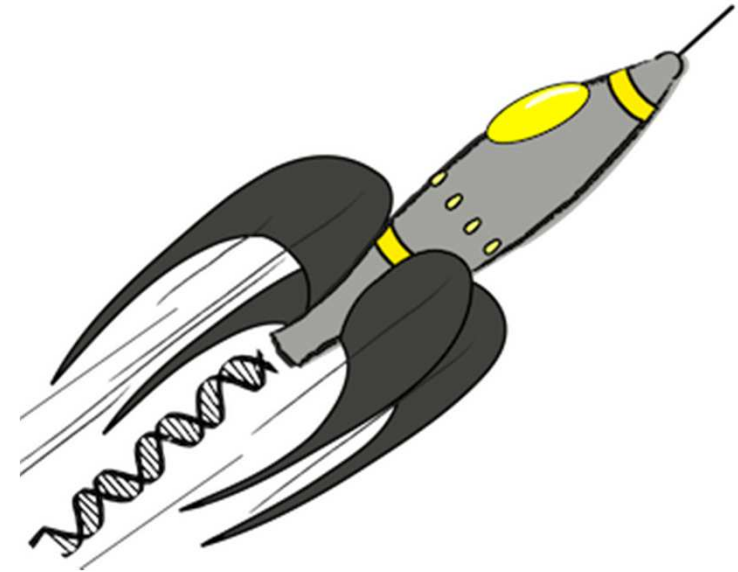
### 3<sup>rd</sup>-partyツール

ツール名	URL
FLASH	<a href="http://ccb.jhu.edu/software/FLASH/">http://ccb.jhu.edu/software/FLASH/</a>
Panda-seq	<a href="https://github.com/neufeld/pandaseq">https://github.com/neufeld/pandaseq</a>
Seq-Prep	<a href="https://github.com/jstjohn/SeqPrep">https://github.com/jstjohn/SeqPrep</a>
PEAR	<a href="http://sco.h-its.org/exelixis/web/software/pear/">http://sco.h-its.org/exelixis/web/software/pear/</a>
FASTQ-Join	<a href="https://code.google.com/p/ea-utils/wiki/FastqJoin">https://code.google.com/p/ea-utils/wiki/FastqJoin</a>

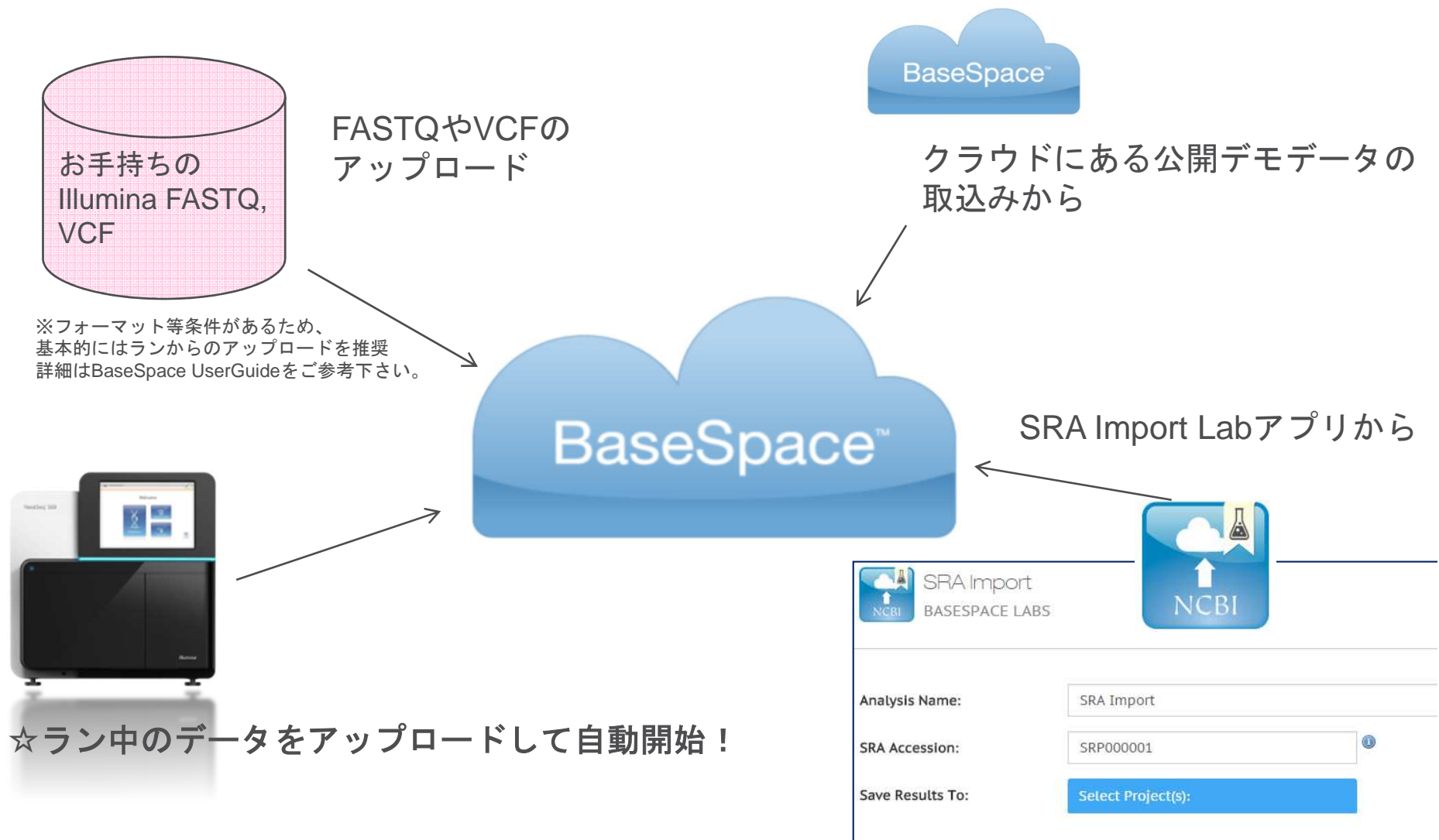
等

## 本日の内容

- イン트로ダクション
- アダプタートリミング
- クオリティトリミング
- ダウンサンプリング
- リードの結合
- 手元のFASTQをトリミングするには

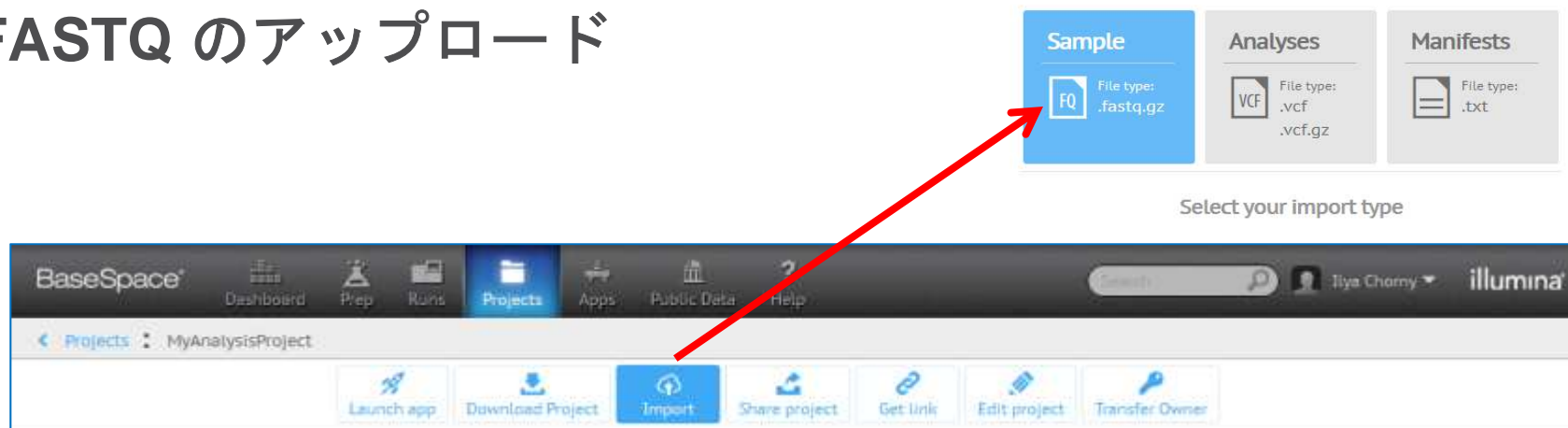


# BaseSpace データ取り込みパターン



※ (SRP\*/ERP\*/DRP\*), experiments (SRX\*/ERX\*/DRX\*), samples (SRS\*/ERS\*/DRS\*), runs (SRR\*/ERR\*/DRR\*), or submissions (SRA\*/ERA\*/DRA\*)対応。ただしイルミナデータのみ、1回のimportは25GBまで。

# FASTQ のアップロード



規約： ☆ イルミナリードのみに対応しており、**ファイル名**が以下のようなイルミナ標準である

SampleName\_SampleNumber\_Lane\_Read\_FlowCellIndex.fastq.gz

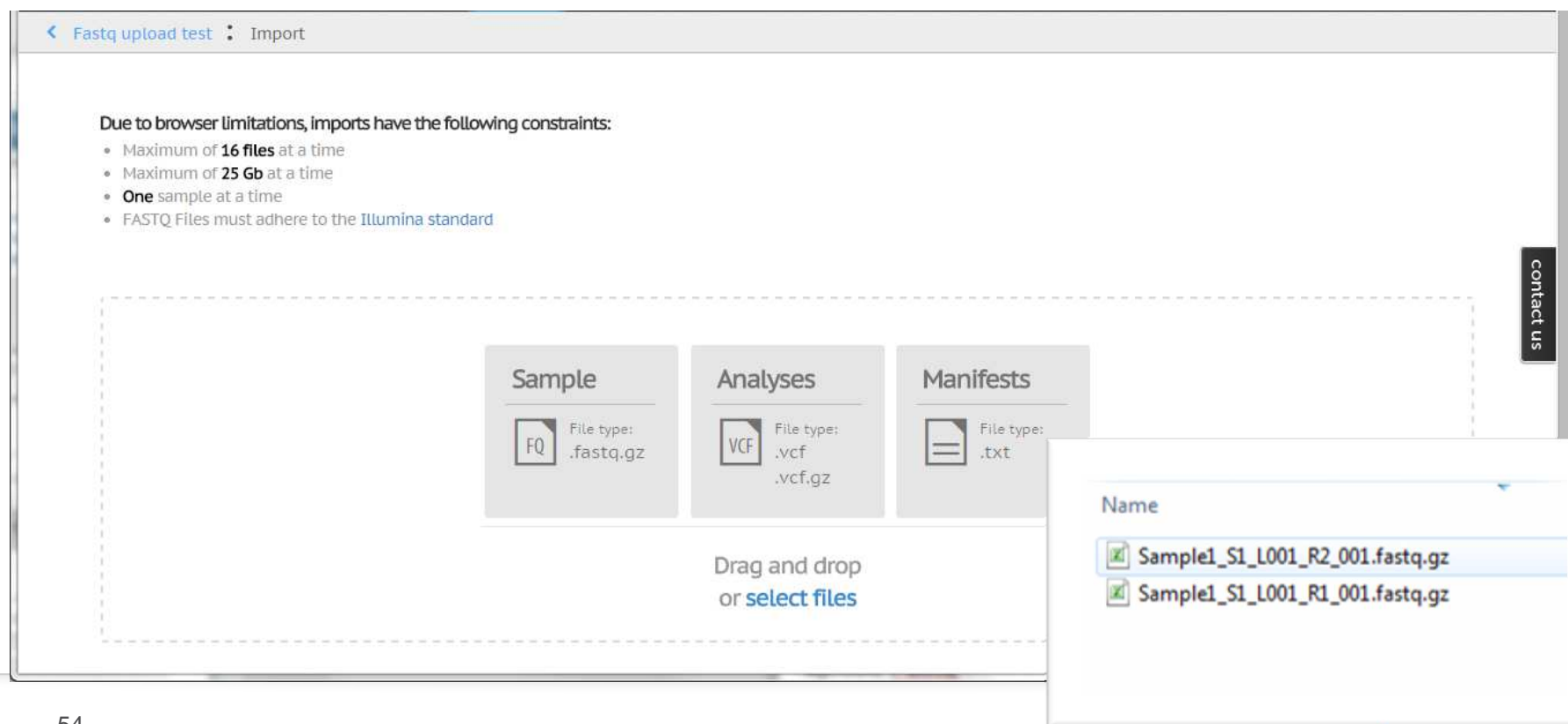
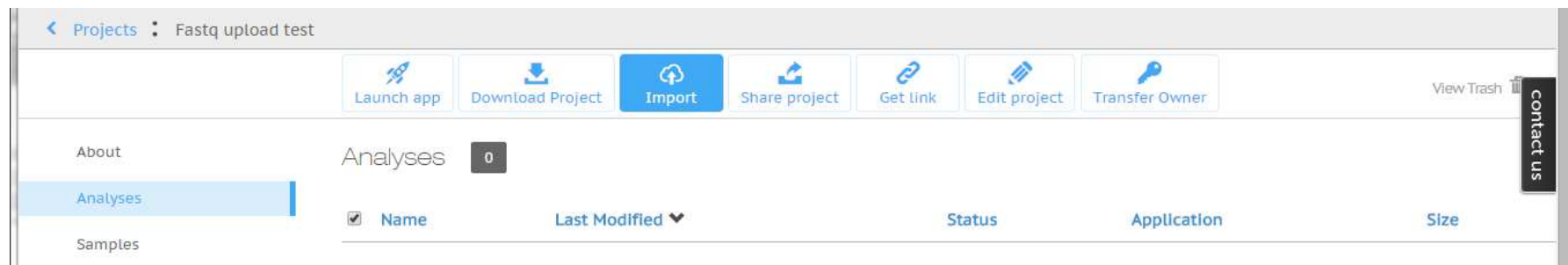
- ☆ **gzip**されている
- ☆ クオリティスコアの数塩基数と一致している
- ☆ **各リードのヘッダ**が以下のようなイルミナ標準を満たしている

@Instrument:RunID:FlowCellID:Lane:Tile:X:Y ReadNum:FilterFlag:0:SampleNumber

ペアードエンドリードの場合さらに；

- ☆ R1とR2でヘッダがペアとして揃ったリード（ReadNumが1と2）が**等数**ある
- ☆ R1, R2ともにPF (Pass Filter)したリード（FilterFlagがN）のみ
- ☆ インポート可能な最大サイズは**25GByte**まで
- ☆ **最大16ファイル/サンプル**
- ☆ **1サンプル単位**で逐次インポート（\* Completely になってから次の処理を開始下さい）

# FASTQ のアップロード



# FASTQ のアップロード

The screenshot shows the BaseSpace web interface for importing FASTQ files. The browser address bar displays <https://basespace.illumina.com/projects/16424410/import>. The navigation bar includes links for Dashboard, Prep, Runs, Projects (active), Apps, Public Data, and Help. The user is logged in as Gabriel Kolle.

The main content area is titled "Fastq upload test : Import". Under "Import type:", the "Sample" option is selected. The sample name is "Sample1" and the reference is "Homo Sapiens - UCSC (hg19)".

Two files are being uploaded:

- `Sample1_S1_L001_R1_001.fastq.gz` (231.70 KB) - 100% complete (green bar with checkmark)
- `Sample1_S1_L001_R2_001.fastq.gz` (253.25 KB) - 50% complete (blue bar)

A purple box highlights the "Complete Import" button. Below the button, a summary of the upload statistics is shown:


Read 1:	151
Read 2:	151
Paired End:	Yes
Reads Passing Filter	100.00%
Total Reads	5,000
Total File Size	484.94 KB

The "Illumina FASTQ file standard" link is also visible.

On the right, there is a "Drag and drop or select files" area with a file type filter set to ".fastq.gz". A "contact us" button is located on the far right.

完了したら、Completeを押下

# FASTQ Toolkit の開始画面から、先ほどアップロードした FASTQ を Select Sample(s): から選択し、トリミングを開始



FASTQ Toolkit v2.0.0  
BaseSpace Labs

Analysis Name: FASTQ Toolkit\_09/04/2015 12:36:21

Input Sample(s) to Process: **Select Sample(s):**

Save Processed Sample(s) to: **Select Project(s):**

Add this string to the output sample name(s):

Default Settings

Minimum read length: 32

The image shows the FASTQ Toolkit v2.0.0 interface. A dashed orange box highlights the 'Select Sample(s):' button under 'Input Sample(s) to Process:' and the 'Select Project(s):' button under 'Save Processed Sample(s) to:'. The 'Analysis Name' field contains 'FASTQ Toolkit\_09/04/2015 12:36:21'. The 'Minimum read length' is set to 32.



ご参考 ;

### **Adapter trimming sequences テクニカルブルテン**

[https://my.illumina.com/MyIllumina/Bulletin/qFYNf9hn\\_kW5SyEZwGOUrA/adapter-sequences-for-use-with-casava-or-bcl2fastq](https://my.illumina.com/MyIllumina/Bulletin/qFYNf9hn_kW5SyEZwGOUrA/adapter-sequences-for-use-with-casava-or-bcl2fastq)

### **Nextera メイトペアのアダプタートリミング**

[http://res.illumina.com/documents/products/technotes/technote\\_nextera\\_matepair\\_data\\_processing.pdf](http://res.illumina.com/documents/products/technotes/technote_nextera_matepair_data_processing.pdf)

### **MiSeq Reporter GenerateFastq ワークフローガイド**

[http://support.illumina.com/content/dam/illumina-support/documents/documentation/software\\_documentation/miseqreporter/miseq-reporter-generatefastq-workflow-guide-15042322-b.pdf](http://support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/miseqreporter/miseq-reporter-generatefastq-workflow-guide-15042322-b.pdf)

### **bcl2fastq 変換ソフトウェア:**

[http://support.illumina.com/downloads/bcl2fastq\\_conversion\\_software.html](http://support.illumina.com/downloads/bcl2fastq_conversion_software.html)

## ご参考；

BaseSpace  
basespace.com

### BaseSpace Fastq Toolkit:

- App について: <http://www.illumina.com/informatics/research/sequencing-data-analysis-management/basespace/basespace-apps/fastq-toolkit-962962.html>
- 紹介ブログ: <http://blog.basespace.illumina.com/2014/12/22/rounding-out-2014-with-new-apps-for-the-basespace-platform-2/>
- サポートアドレス: [basespacelabs@illumina.com](mailto:basespacelabs@illumina.com)

BaseSpaceコアアプリ各ワークフローのフローチャート図は各ユーザガイドにあります  
[support.illumina.com/downloads/basespace\\_core\\_apps\\_user\\_guides.html](http://support.illumina.com/downloads/basespace_core_apps_user_guides.html)

BaseSpace最新News  
blog.basespace.illumina.com      #RSS 購読可能

ヘルプセンター（ウェブヘルプ）  
help.basespace.illumina.com

サポートウェビナーにご参加いただき  
ありがとうございました。



本日のセッション終了後のご質問は、  
[techsupport@illumina.com](mailto:techsupport@illumina.com)  
で承ります。