

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  
発生川絵里 (Eri Kibukawa)

※BaseSpace アプリ: FASTQ toolkit /smallRNA/ FASTQC

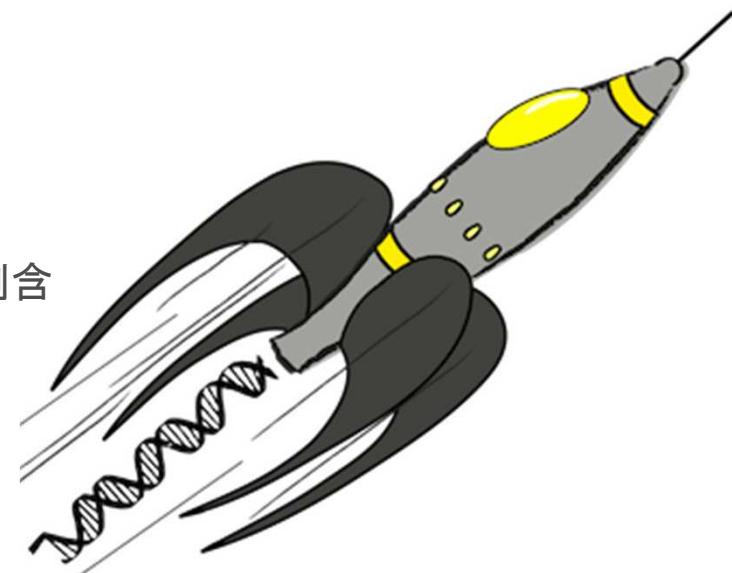


© 2013 Illumina, Inc. All rights reserved.  
Illumina, IlluminaDx, BaseSpace, BeadArray, BeadXpress, cBot, CSPro, DASL, DesignStudio, Eco, GAIIx, Genetic Energy, Genome Analyzer, GenomeStudio, GoldenGate, HiScan, HiSeq, Infinium, iSelect, MiSeq, Nextera, NuPCR, SeqMonitor, Solexa, TruSeq, TruSight, VeraCode, the pumpkin orange color, and the Genetic Energy streaming bases design are trademarks or registered trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners.

illumina®

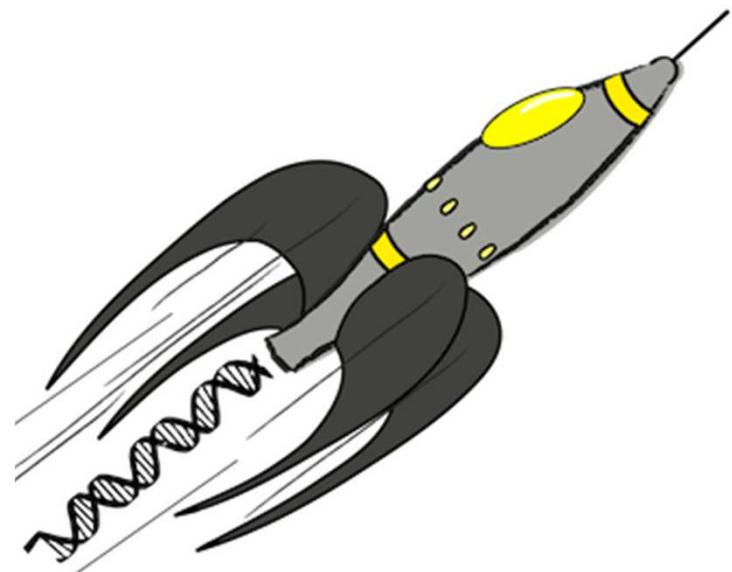
# 本日の内容

- イントロダクション
- 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

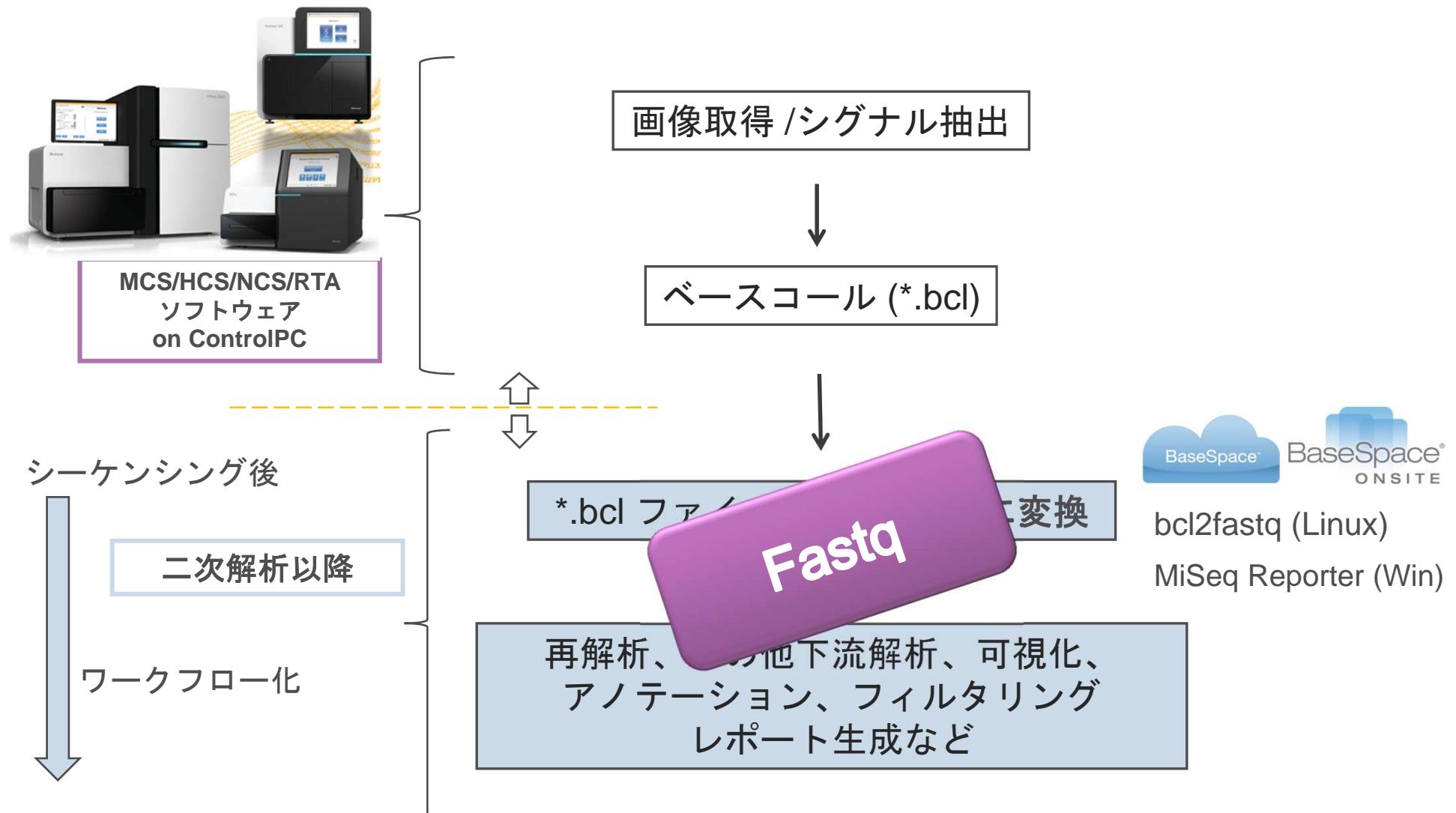


# 本日の内容

- イントロダクション
- 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:CTTGT  
Sequence ➔ TGAAACCAGTGGTCTTAATTGGCATTACACACACACACAGAATTAAAAAAAATCAAAGG  
+  
Q-score ➔ =55>7 ;?::BDADDD@EE88DCD?DFFEFFECBE6666BB=B;<;<-34:;<CB51>=BBEE>EE?  
➔ @HWI-BRUNOP20X:994:B809UWABXX:1:1101:13660:2247 1:N:0:CTTGT  
➔ CCAAACATTAAGTAACTCTTAAATGGCACACAGGTTTAAAGCTATTGGTTTCCTCCTAACT  
+  
➔ FFEDFBGECCCCDFGEFFFFGGDF=FBFFFGGGE7CEEDEFBFBFGEEGF@FCDDFDFFEGFEAGF  
➔ @HWI-BRUNOP20X:994:B809UWABXX:1:1101:13966:2183 1:N:0:CTTGT  
➔ TTGGGTAACCTGAATATAACATGGCTCCCTGCTGTAAGCAAATGTTAGAGCTGAATTTTCCT  
+  
➔ HHHHHHEHHHHHHFHHHHHHHHHHHHHHHHGGFHHHHHHHHHFHHHFHEHHFHEHHHFFHHHF

# FASTQの生成場所・方法



**MiSeq**



**MiSeq  
Reporter**

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



**NextSeq**



**HiSeq**



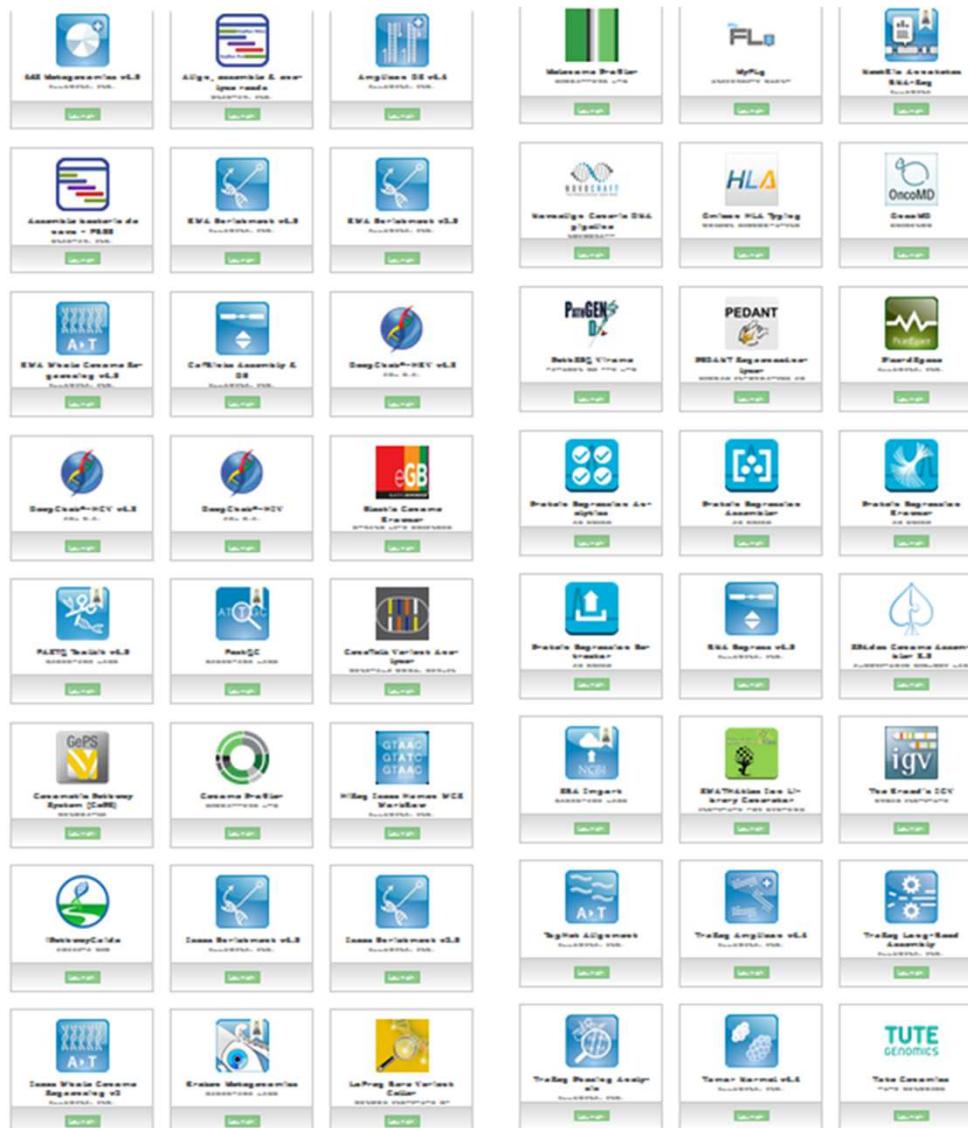
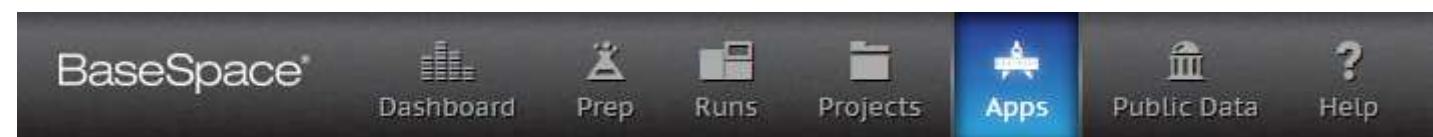
A diagonal band of sequence data is shown, consisting of multiple lines of blue text representing DNA sequence reads.

お使いのLinux server

**bcl2fastq2**

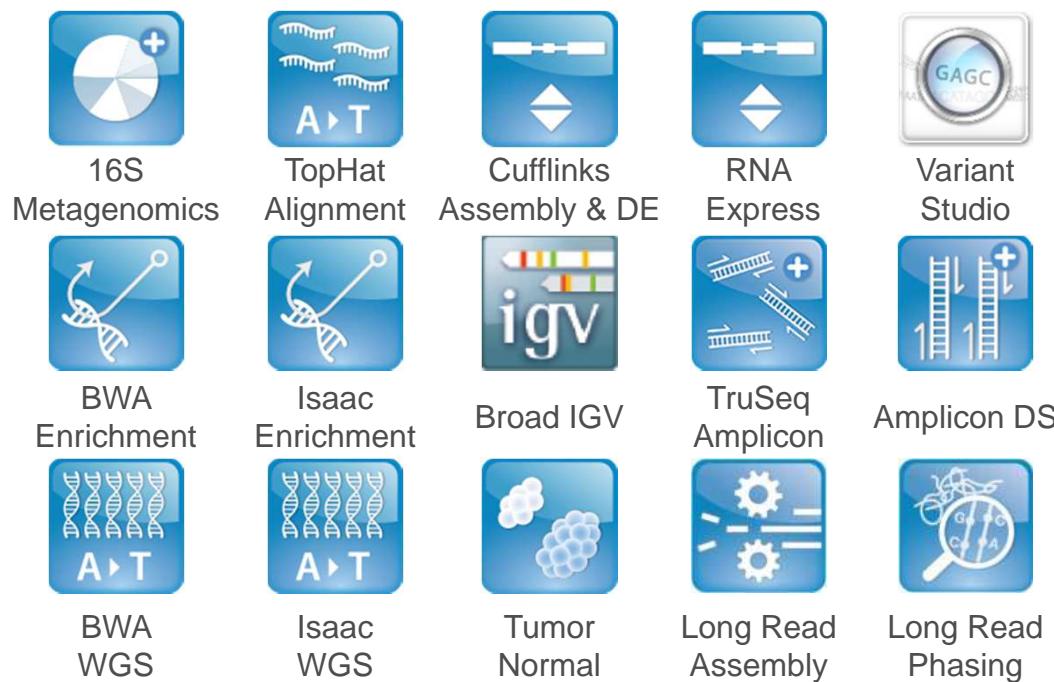


# アプリ (>60)

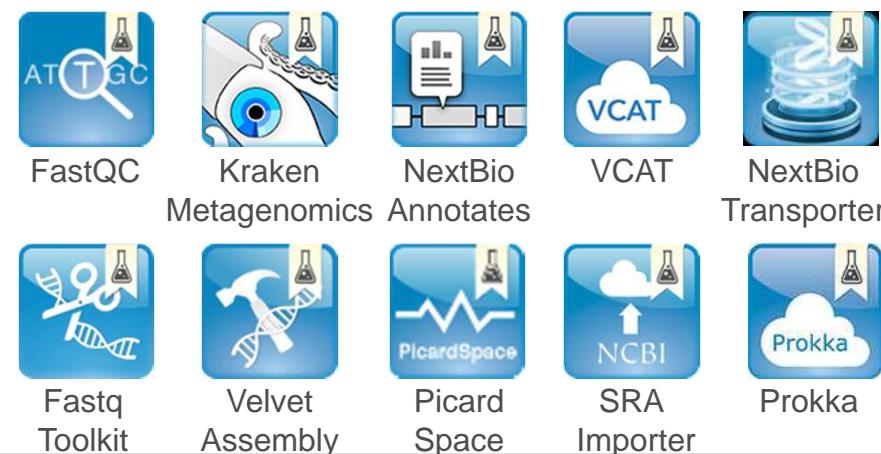


| Categories           |                         |
|----------------------|-------------------------|
| Resequencing         | Small RNA               |
| Targeted Sequencing  | De Novo Assembly        |
| RNA-Seq              | Gene Fusion Detection   |
| ChIP-Seq             | Methyl-Seq          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

## <イルミナコアアプリ>

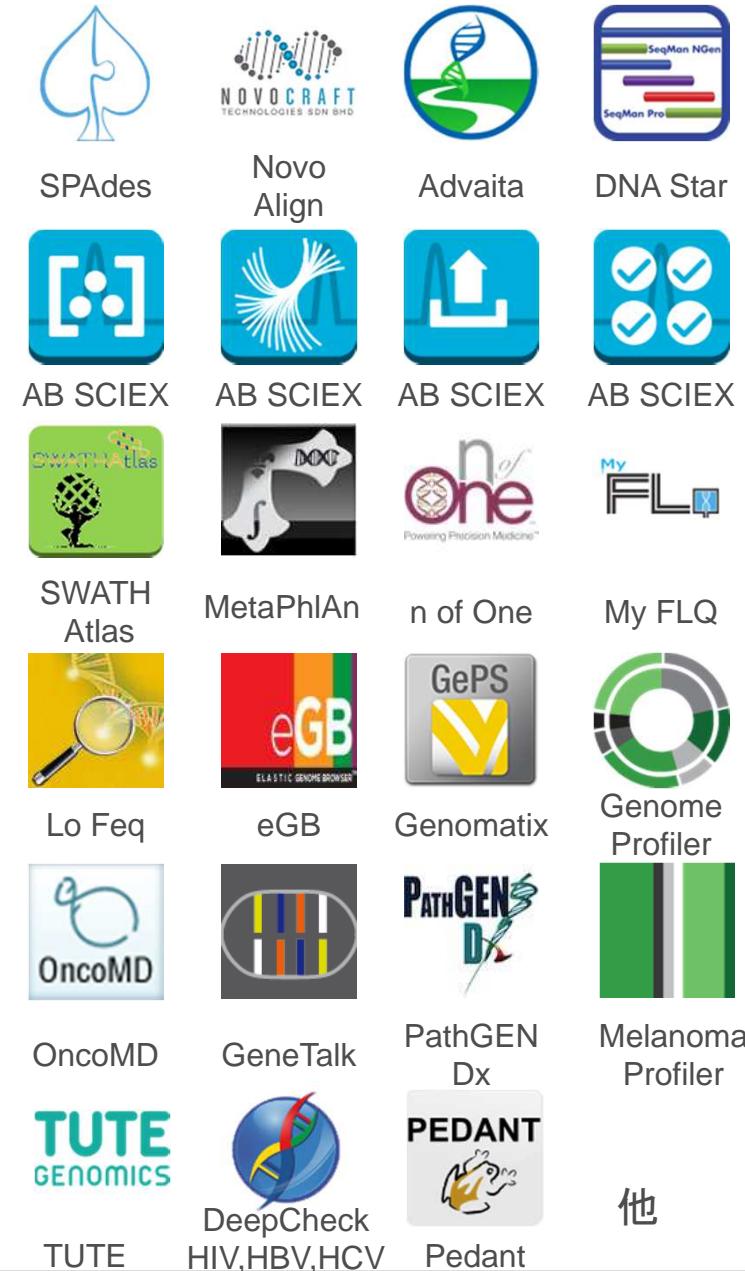


## <イルミナラボアプリ>



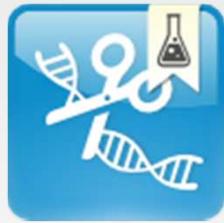
他

## <他社製アプリ>



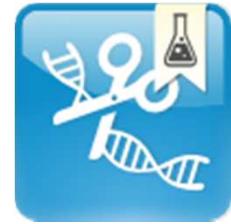
# BaseSpace Labsアプリ (準サポート)

- ▶ 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)).

|  |   |  |   |   |
|--|---|--|---|---|
| <b>FASTQ Toolkit</b><br> <ul style="list-style-type: none"><li>▶ Sub-sample reads</li><li>▶ Trim Adapters</li><li>▶ Trim Bases</li><li>▶ Ploy A/T trimming</li><li>▶ Quality Trimming</li><li>▶ Read Filtering</li><li>▶ Reverse Complement</li></ul> | <b>FastQC</b><br> <ul style="list-style-type: none"><li>▶ Perform QC of raw sequencing data.</li><li>▶ Determine adapter contamination</li></ul> | <b>VCAT v2.3</b><br> <ul style="list-style-type: none"><li>▶ Compare Variant Call Sets to standards</li><li>▶ Intersect variant call sets.</li></ul> | <b>SRA Import v0.0.3</b><br> <ul style="list-style-type: none"><li>▶ Import up to 25GB of sequencing data from SRA</li></ul> | <b>SRA Submission v0.0.3</b><br> <ul style="list-style-type: none"><li>▶ Deposit sequencing data in SRA.</li></ul> |
|--|---|--|---|---|

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 (クオリティートリミング)

3'-端の低クオリティー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

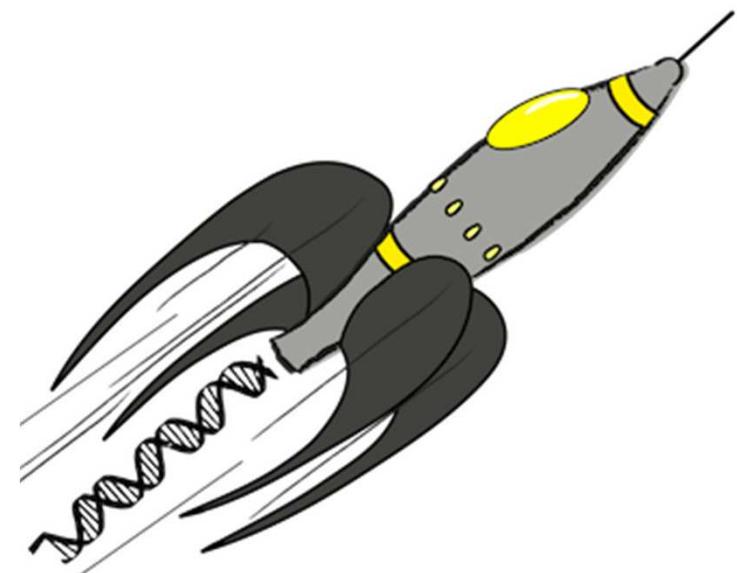
(Nexteraメイトペアリードからペアードエンドリード 方向への変換目的など) に加え、  
他ペアードエンドリードが 1 つのFASTQからR1, R2への振り分け

## 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

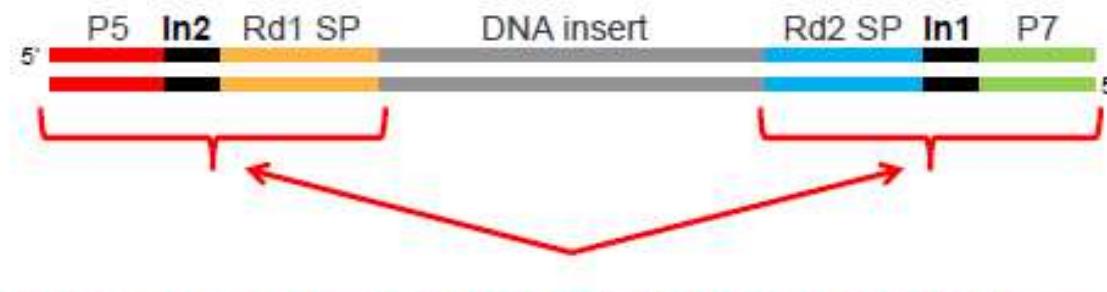
# 本日の内容

- イントロダクション
- 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）

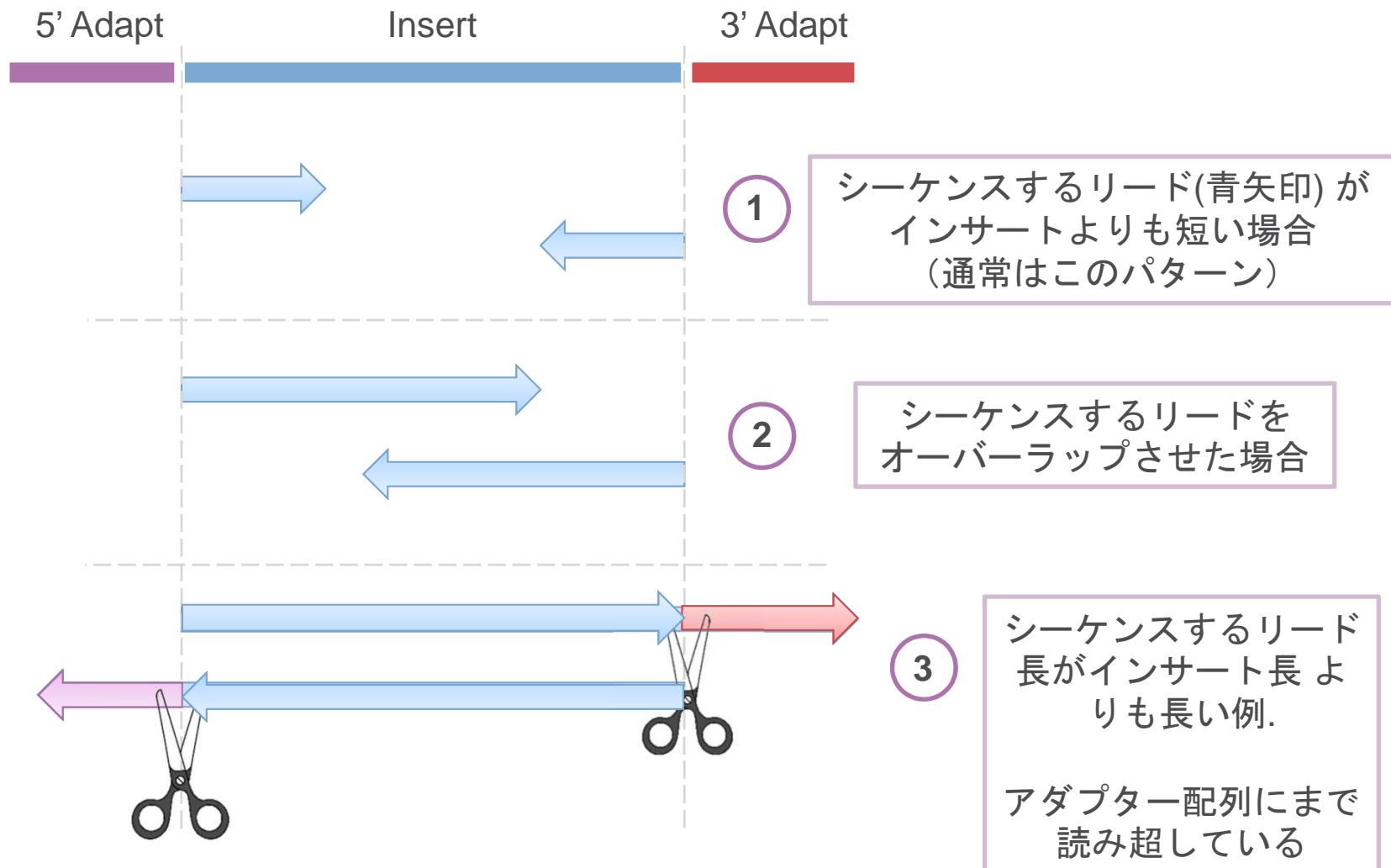
**ライブラリ = DNA インサート + 両端にそれぞれ別のアダプター**

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

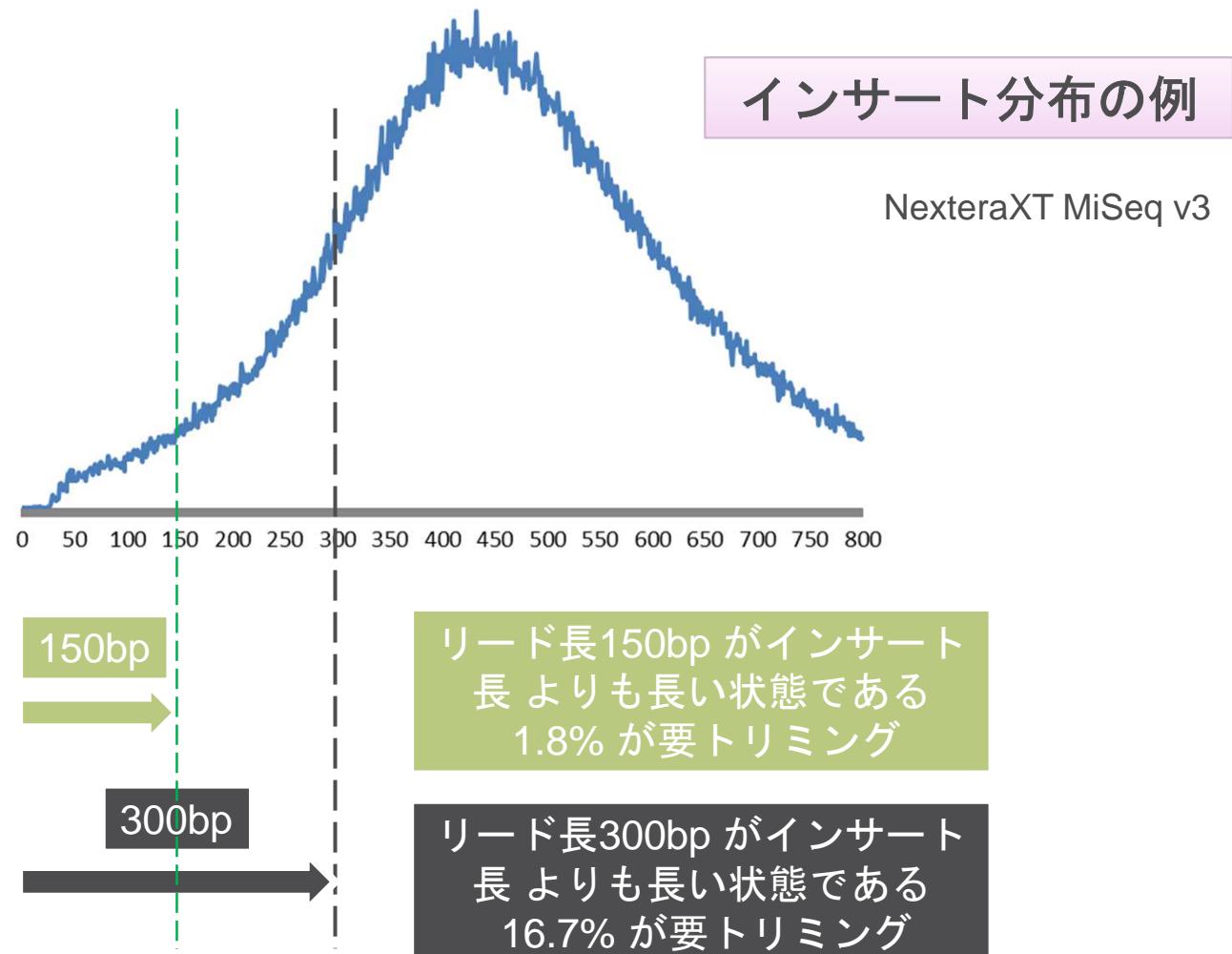
※ 詳しくは、弊社サポートウェビナー 2015/07/10 をご参考いただけます。  
SBS (Sequencing By Synthesis) ケミストリーとは何か？  
[http://www.illuminakk.co.jp/events/webinar\\_japan/support\\_webinar.ilmn](http://www.illuminakk.co.jp/events/webinar_japan/support_webinar.ilmn)

# インサート長とアダプタートリミング

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         |   |
| [Reads]           |   |
|                   | 151   |
|                   | 151   |
| [Settings]        |   |
| Adapter           | AGATCGGAAGAGCACACGTCTGAACCTCCAGTCA              |
| AdapterRead2      | AGATCGGAAGAGCGTCGTAGGGAAAGAGTGT                 |
| [Data]            |   |
| Sample_ID         | Sample_N Sample_P Sample_V Sample_P Description |
| test              | test test                                       |

### 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)

# 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

```
@M00000:71:00000000-D00LW:1:1101:16265:1658 1:N:0:1
ACTCTGCGTTGCGCTTCTGCTCGGCCTCCAGCTCACCCCTCCCTGTCTCTTATACACATCTCCGAGCCCA
+
BCCCCFFCCBCCGGGGGGGGGGGGGGHHHHHHHHHHHHHHGGHHHHHHHHHHHHHHGGGGGH
```

## アフター

```
@M00000:71:00000000-D00LW:1:1101:16265:1658 1:N:0:1
ACTCTGCGTTGCGCTTCTGCTCGGCCTCCAGCTCACCCCTCC
+
BCCCCFFCCBCCGGGGGGGGGGGGHHHHHHHHHHHHHG
```



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           | AGATCGC       |
| AdapterRead2      | AGATCGC       |
| [Data]            |               |
| Sample_ID         | Sample_N      |
| 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

## アフター

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、  
クオリティースコアは一律2(＃)で置換

# BaseSpaceでトリミング目的に使えるツール

## FASTQ Toolkit

The screenshot shows the BaseSpace FASTQ Toolkit interface. Key elements include:

- Analysis Name:** A text input field.
- Input Sample(s) to Process:** A red-bordered text input field.
- Select Sample(s):** A blue button.
- Save Processed Sample(s) to:** A red-bordered text input field.
- Select Project(s):** A blue button.
- Add this string to the output sample name(s):** A text input field.
- Sub-sampling**, **Adapter Trimming**, **Base Trimming**, **Poly-A/T Trimming**, **Quality Trimming**, **Read Filtering**, **Modify Reads**: A list of processing steps.
- BaseSpace Labs:** A section with a checkbox and a detailed text about usage terms.
- Adapter Trimming:** A section with a description and a dropdown menu.
- Adapter trim stringency (0.01-0.99):** A numeric input field set to 0.9.
- Select an adapter to trim or specify the adapter sequence(s) below:** A dropdown menu showing "None selected".
- Adapter sequence(s) to trim from the 5'-end:** An input field.
- Adapter sequence(s) to trim from the 3'-end:** An input field containing `CTGTCTCTTATAACACATCTCCGAG`. This field is highlighted with an orange rectangle.
- Trim Ns from the 3'-end before identifying adapters:** A dropdown menu.

## 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><br>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?



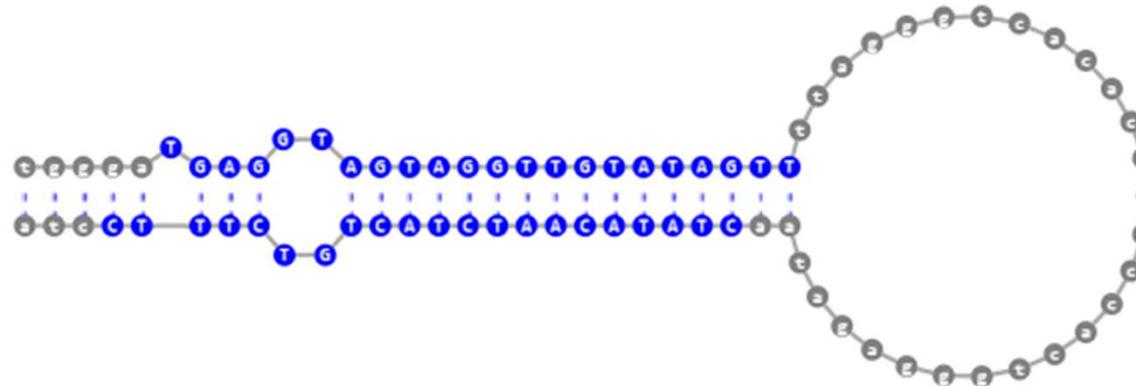
Small RNA v1.0

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ートリム済みとなる。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、アダプタートリム済みのため、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



BaseSpace®

AGCGATGTCCTCAAATGATTAGACCACTTACCAATTG  
AGATAACATCACAAAGGTTACCCACAATTG  
TGCAATGGTAGATACAGTAGCTAACACATTA  
AATTAAATCTATCAAAGGGAACAGGTTACAA  
GACCGTTATGAAATTATAGCTATGAGCTTAA  
TAGATATAGCAAAGAGTCACCTTCTT  
TAGTGATTAGCAGGTGATCTT  
CTTTAAAGAGCTTAAATGCA  
GACCACGATACGATATCAAGACTAC  
GAGGAGGACCTACAGTGTACAG  
AGCCAAACCAATTG

- 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 sequences(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

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:

TGGAATTCTCGGGTGCCAAC

Trim Ns from the 3' end before identifying adapters:

ドロップダウンから選べるキットもある

# 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

The screenshot shows a software interface for configuring a BaseSpace Labs App. At the top, there are three input fields: 'Maximum GC content', 'Minimum sequence complexity', and 'Only keep reads passing filters'. Below these is a section titled 'BaseSpace Labs:' containing a text area and a checkbox. A large purple oval highlights the checkbox and its associated text. An arrow points from the Japanese instruction 'AS-ISでご使用いただくことの明示的ご了承' to the highlighted checkbox.

MAXIMUM GC content:

Minimum sequence complexity:

Only keep reads passing filters:

BaseSpace Labs:

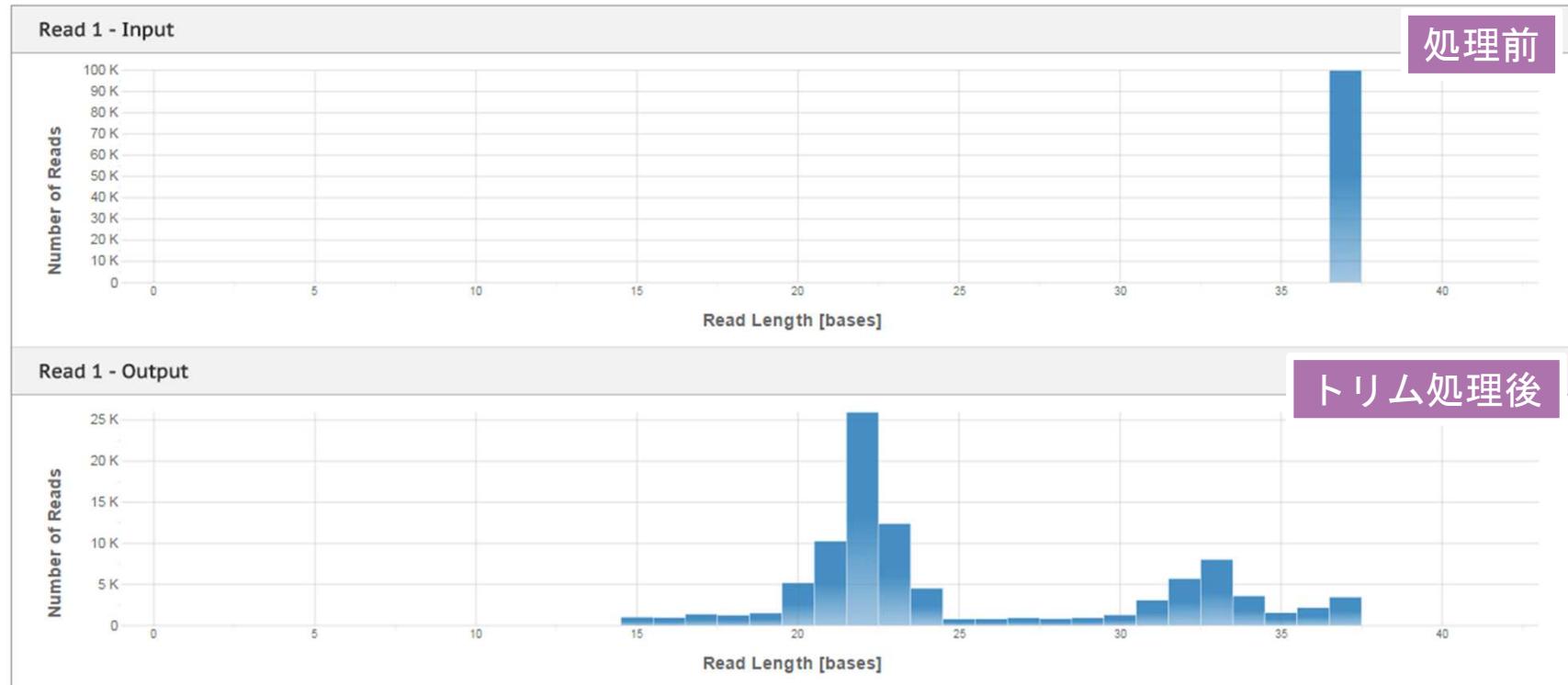
AS-ISでご使用いただくことの明示的ご了承

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.

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

The screenshot shows the Small RNA v1.0 application interface. At the top left is a logo with a blue square containing a white RNA hairpin and a green circle with a plus sign. To its right is the text "Small RNA v1.0" and "ILLUMINA". Below this is a green "Launch" button. The main area has a light gray background. At the top center is a search bar labeled "Precursor: hsa-let-7a-1". Below it is a diagram of the precursor RNA structure, which is a linear strand with a large loop on the right. A "Save as PNG" button is located below the diagram. To the left of the diagram, the text "hsa-let-7a-1 in miRBase" is displayed, followed by "5' Counts: 203771", "3' Counts: 129", and "5'/3' ratio: 1579.62". Below this is a sequence alignment table:

| Sequence  | Count  |
|---|--------|
| tgggaTGAGGTAGTAGTTGTATAGTtttagggtcacaccaccactggagataaCTATAACATCTACTGTCCTTCcta |        |
| (((((.....(((((((((.....)))))))))))).....)))))))))))                          | 6      |
| ....TCTGAGGTAGTAGTTGTATAGTT.....  | 196698 |
| ....TGAGGTAGTAGTTGTATAGT.....   | 7061   |
| ....TGAGGTAGTAGTTGTGTGGTTT.....   | 6      |
| .....CTATAACATCTACTGTCCTTC...33   |        |

At the bottom of the main window are three small thumbnail images representing different features of the application.

**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 miRNAs, 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.

**Categories**

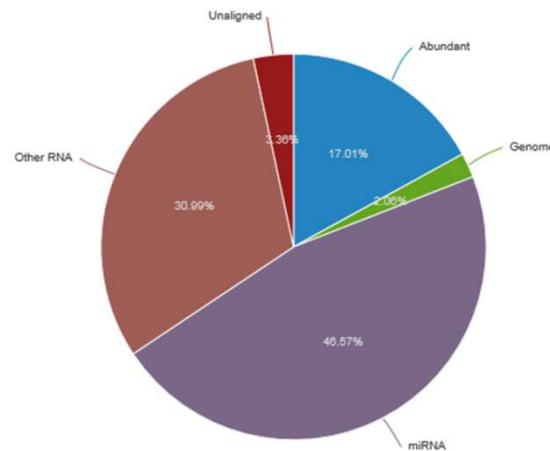
- Native
- Small RNA
- RNA-Seq

**Differential Expression**

**Support**



# Small RNA アプリ結果のレポート



| Hsa-mirna-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 is defined 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 and sequencing.

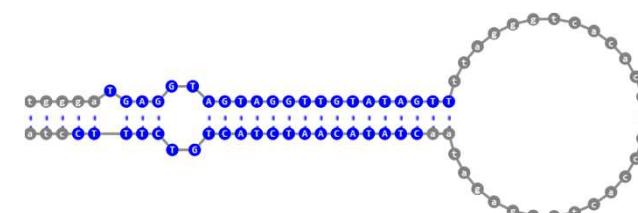
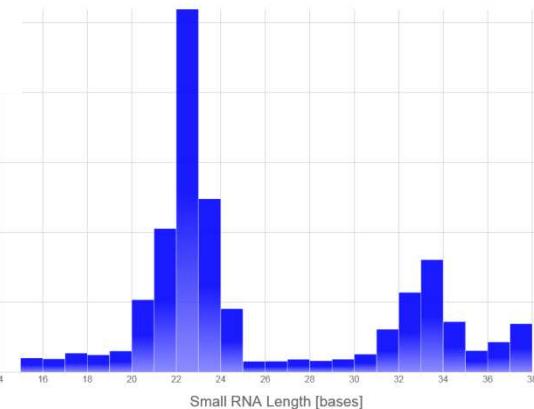
| Number of IsomiRs (Known Precursor) with Reads | Total Number of Reads |
|--|-----------------------|
| 464  | 22,967                |

## Length Distributions

Reads are listed for each type. If there are less than 10 sequences with reads, 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                |



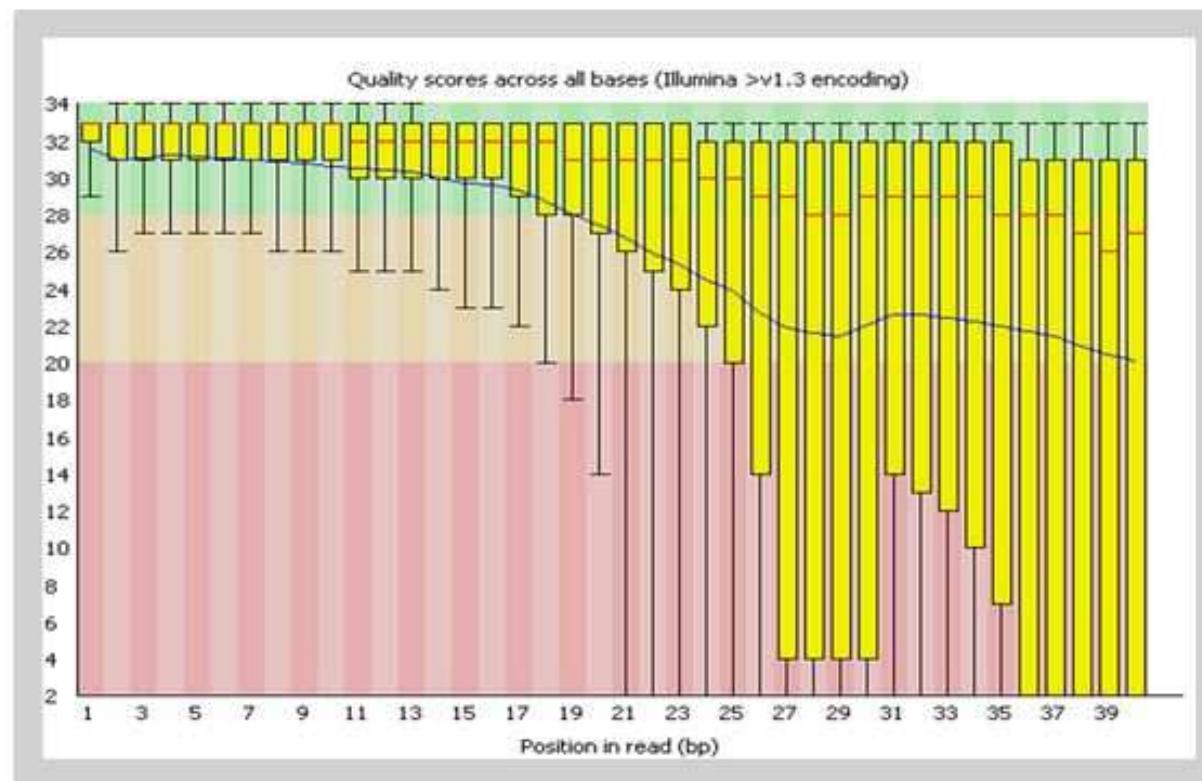
# Small RNA

このFASTQリードはトリムされたものか？  
– FastQCアプリ



FastQC  
BASESPACE LABS

Launch



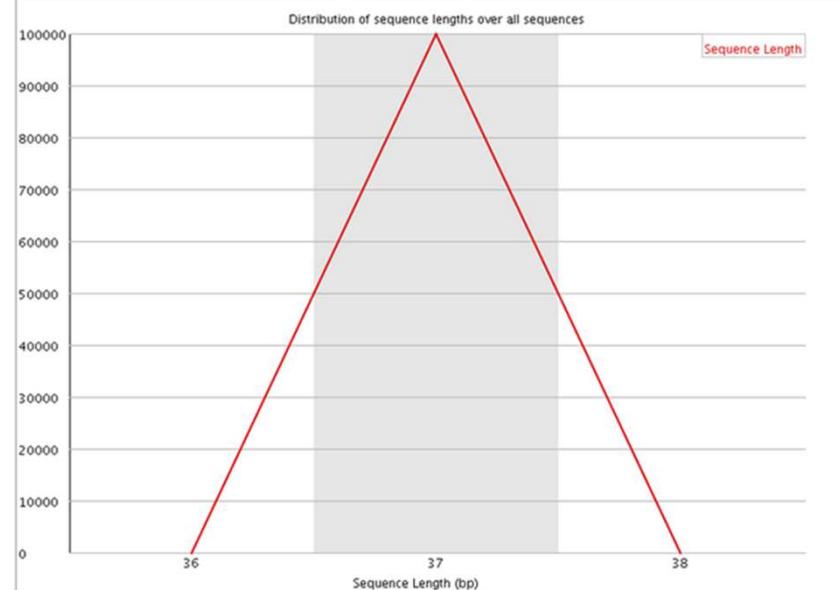
# Small RNA

このFASTQリードはトリムされたものか？  
– FastQCアプリ



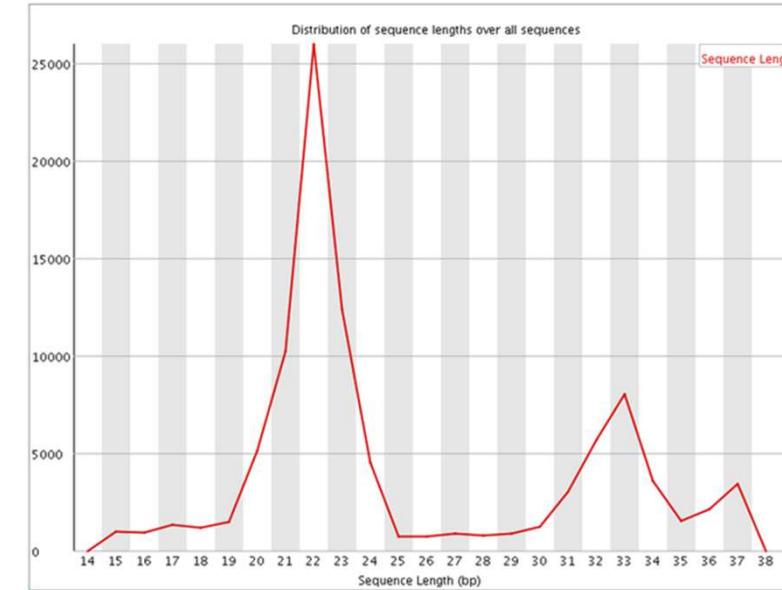
トリムされていない

✓ Sequence Length Distribution



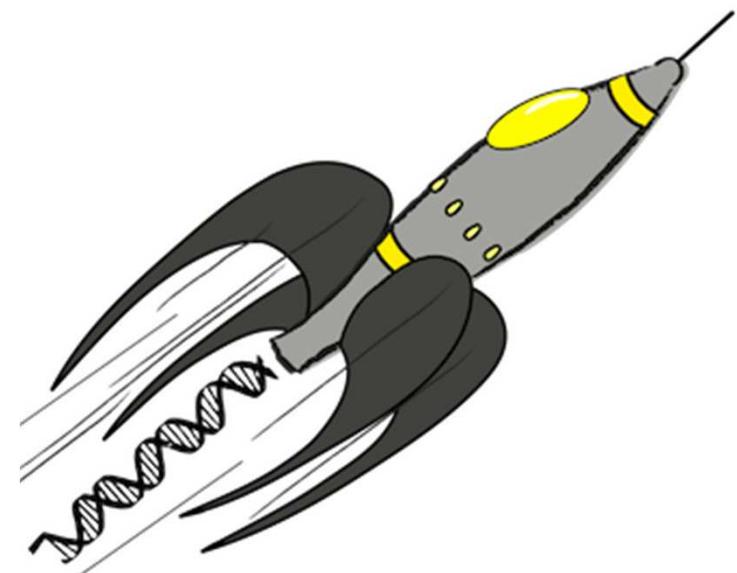
トリムされている

⚠ Sequence Length Distribution



# 本日の内容

- イントロダクション
- 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'末端のクオリティー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イ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ーが閾値を下回ったときに以降をトリムするものが多い
- どういう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           | AGATCGGAAGAGCACACGTCTGAACTCCAGTCACG |              |
| Adapter Read2     | AGATCGGAAGAGCCTCGTGTAGGGATGTTTGATG  |              |
| QualityScoreTrim  | 20                                  |              |
| [Data]            |                                     |              |
| Sample_ID         | Sample_Name                         | Sample_Plate |
| test              | test                                | test         |

**QualityScoreTrim**

**[settings]**  
**QualityScoreTrim, <qualityScore>**

# QualityScoreTrim

[settings]  
QualityScoreTrim,<qualityScore>

# Qスコアによるトリミングの例

QualityScoreTrim,20

ビフォー

```
@M00000:72:00000000-D00LW:1:1101:22420:18334 1:N:0:1  
CACCAAGGGCCTGGGTGTCAATGGCGGGCTTGTGACTGCACAAAAGGGCCTCCGCAGGGCTCCGCC  
+  
BBBBBBFBBBBBGGGEEFGGGHHHGGG00>10B355@BB3@3BG1?E1///1B11//////////?///
```



アフター

```
@M00000:72:00000000-D00LW:1:1101:22420:18334 1:N:0:1  
CACCAAGGGCCTGGGTGTCAATGGCGGGCTTGTGACTGCACAAAAGG  
+  
BBBBBBFBBBBBGGGEEFGGGHHHGGG00>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

The screenshot shows the BaseSpace FASTQ Toolkit application interface. At the top left is the 'BaseSpace' logo and a back arrow. To the right is the 'illumina' logo. Below the header, there's a blue icon of a DNA helix and the text 'FASTQ Toolkit' and 'BaseSpace Labs'. The main area has several input fields: 'Analysis Name:' with the value 'FASTQ Toolkit v1.0 03/17/2015 7:07:19', 'Input Sample(s) to Process:' with a 'Select Sample(s)' button, and 'Save Processed Sample(s) to:' with a 'Select Project(s)' button. There's also a field 'Add this string to the output sample name(s:)'. On the right side, a grey box says 'This app is free.' with a large blue arrow pointing right. A vertical bar on the far right says 'contact us'. A dashed orange line points from the 'Quality Trimming' option in the sidebar to the 'Trim the 3'-end of reads with quality level:' input field, which is highlighted with a yellow border. The sidebar also lists other options: Sub-sampling, Adapter Trimming, Base Trimming, Poly-A/T Trimming, Quality Trimming (which is selected), Read Filtering, and Modify Reads. At the bottom, there's a section for 'BaseSpace Labs:' with a checkbox and a detailed disclaimer text.

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

This app is free.

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:

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.

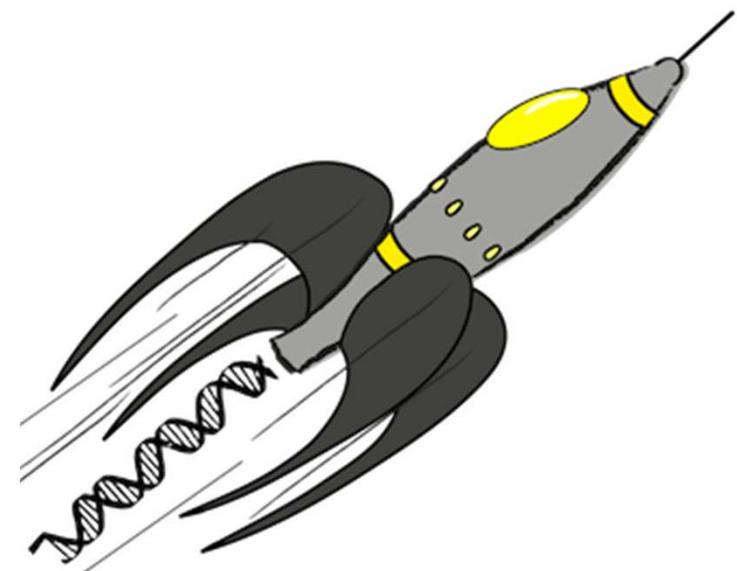
## Quality トリミング 3<sup>rd</sup>- party ツール例

| 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

- イントロダクション
- 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、ディスク領域が足らないなど
- 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

The screenshot shows the BaseSpace FASTQ Toolkit application interface. The top navigation bar includes a back arrow, the title "BaseSpace", and the "illumina" logo. The main header "FASTQ Toolkit" is displayed above the "BaseSpace Labs" section.

**Analysis Name:**  FASTQ Toolkit

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

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

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

**Process Options:**

- ▶ 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 ASL. I understand that ASL has no obligation to support or maintain this software, and that I am using it at my own risk. I also understand that ASL is not liable for any damages or losses that may result from my use of this software, including but not limited to, direct, indirect, incidental, special, punitive, and consequential damages, and that I am solely responsible for any damages or losses that may result from my use of this software.

**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 constrains) 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:**

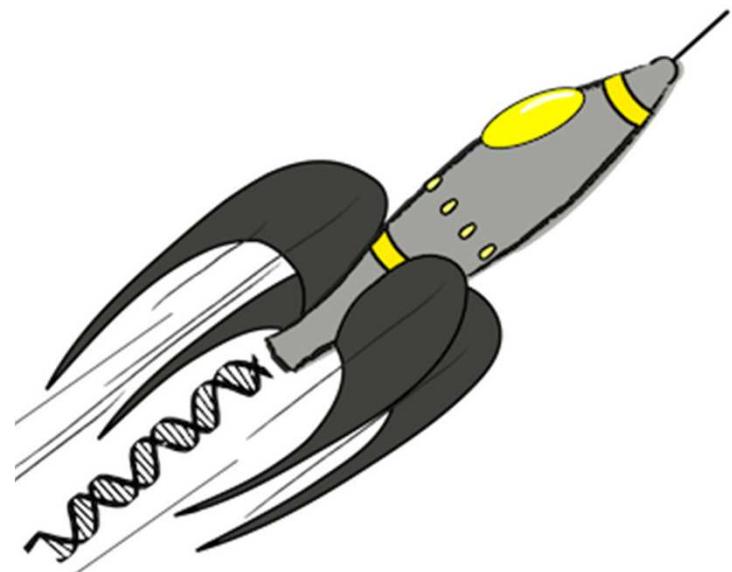
**Maximum number of bases to keep:**

**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

- イントロダクション
- 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

通常はある程度クオリティーの良い塩基のオーバ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)

## ○ どういう時に使うものなのか?

- 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

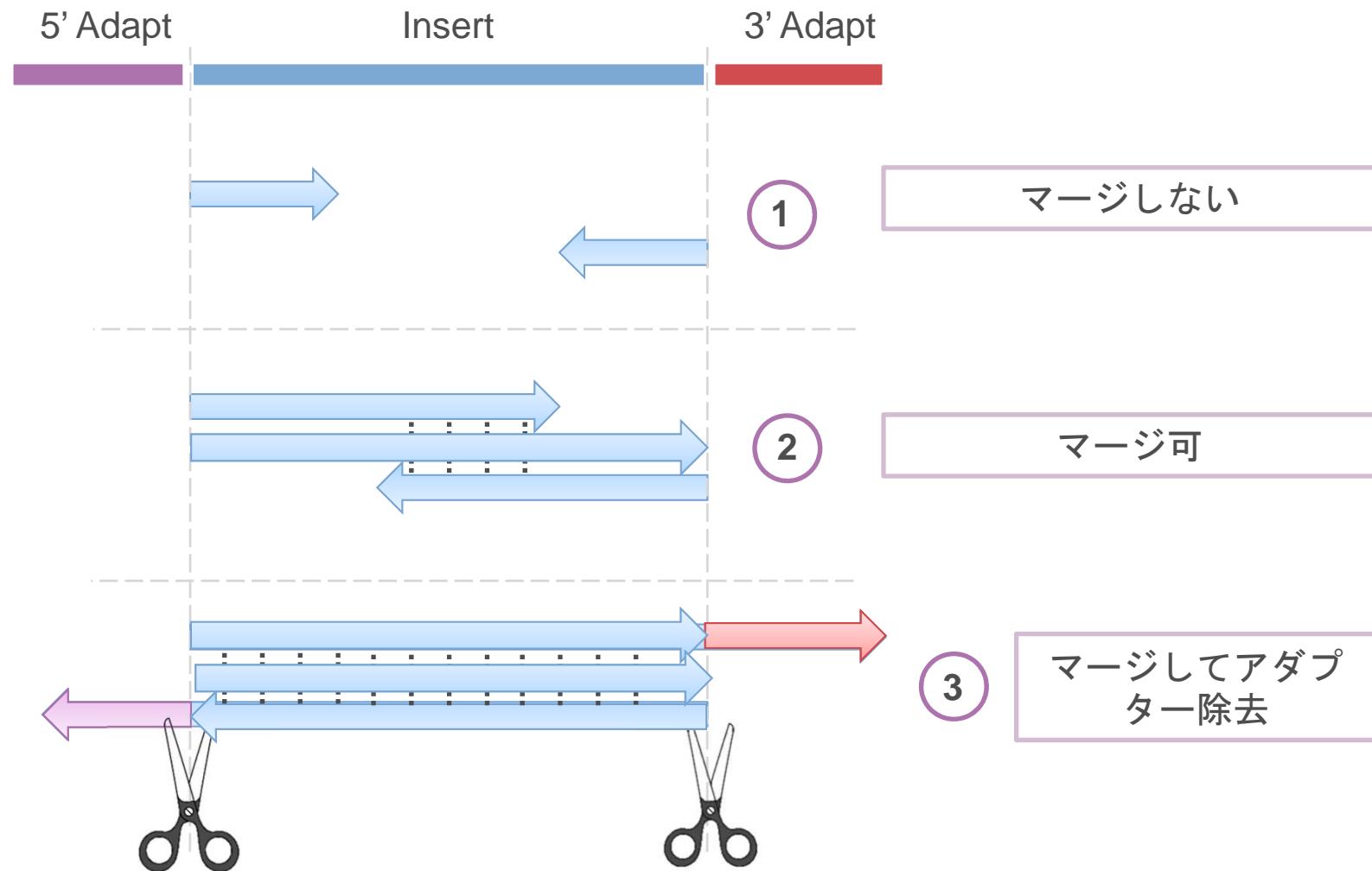
## ○ 逆に、適さないときは?

- クオリティーの良い塩基のオーバラップがない
- 一部のリードしかオーバラップがない場合 (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



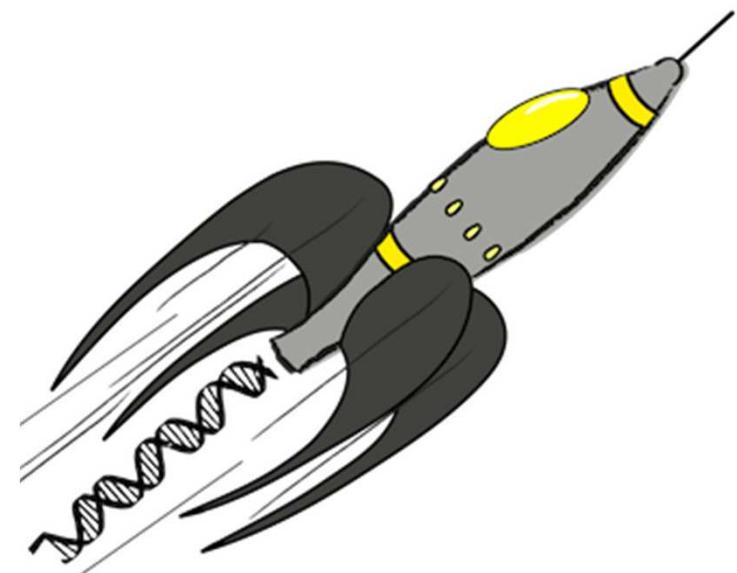
# リードマージができるツールの一例 3<sup>rd</sup>-partyツール

| 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> |

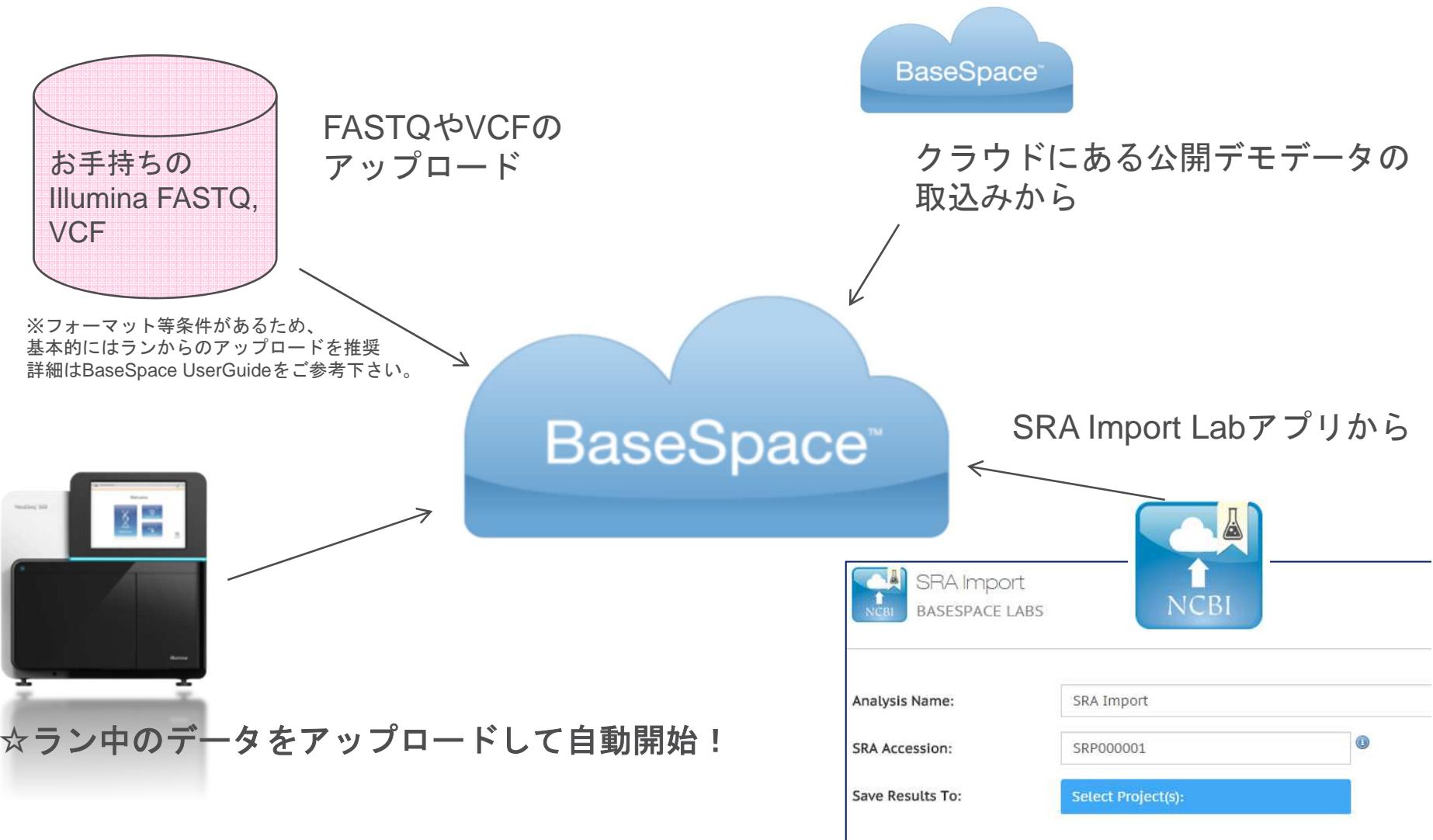
等

## 本日の内容

- イントロダクション
- 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\*)に対応。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

The screenshot shows the BaseSpace software interface. At the top, there's a navigation bar with 'BaseSpace' logo, 'Dashboard', 'Prep', 'Runs', 'Projects' (which is selected), 'Apps', 'Public Data', and 'Help'. Below the navigation bar, the title 'Projects > MyAnalysisProject' is visible. In the center, there's a toolbar with icons for 'Launch app', 'Download Project', 'Import' (which is highlighted in blue), 'Share project', 'Get link', 'Edit project', and 'Transfer Owner'. To the right of the toolbar, a dropdown menu titled 'Select your import type' is open, showing three options: 'Sample' (selected, indicated by a blue background and a red arrow pointing to it), 'Analyses', and 'Manifests'. Each option has a small icon and a description: 'Sample' has an FQ icon and 'File type: .fastq.gz', 'Analyses' has a VCF icon and 'File type: .vcf .vcf.gz', and 'Manifests' has a manifest icon and 'File type: .txt'.

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が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 (\* Completeになってから次の処理を開始下さい)

# FASTQ のアップロード

The screenshot shows a web-based project management interface. At the top, there's a navigation bar with links for 'Launch app', 'Download Project', 'Import' (which is highlighted in blue), 'Share project', 'Get link', 'Edit project', and 'Transfer Owner'. Below the navigation bar, there are three tabs: 'About', 'Analyses' (which is selected and highlighted in blue), and 'Samples'. The 'Analyses' tab displays a table with columns for 'Name', 'Last Modified', 'Status', 'Application', and 'Size'. A small '0' is shown in the 'Analyses' count box. On the right side of the interface, there are two 'contact us' buttons.

The screenshot shows an 'Import' dialog box. It starts with a note: 'Due to browser limitations, imports have the following constraints:' followed by a list:

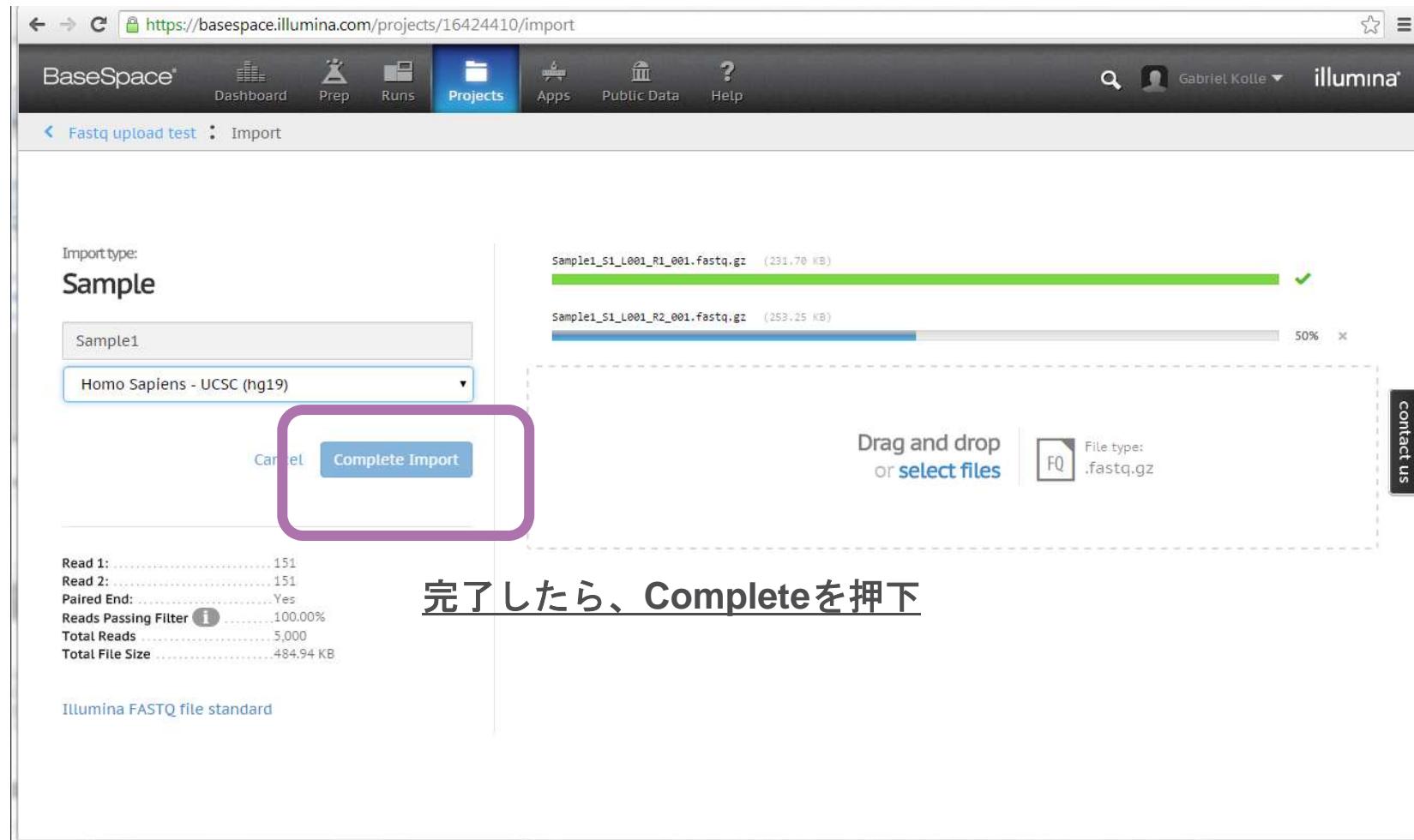
- Maximum of **16 files** at a time
- Maximum of **25 Gb** at a time
- **One sample** at a time
- FASTQ Files must adhere to the [Illumina standard](#)

Below this, there are three categories: 'Sample', 'Analyses', and 'Manifests'. Each category has a file icon and a file type description: 'Sample' (FQ, .fastq.gz), 'Analyses' (VCF, .vcf, .vcf.gz), and 'Manifests' (List icon, .txt). A large dashed rectangular area is provided for dragging and dropping files. Below this area, it says 'Drag and drop or [select files](#)'. To the right, a file selection dialog is open, showing a 'Name' column with two entries:

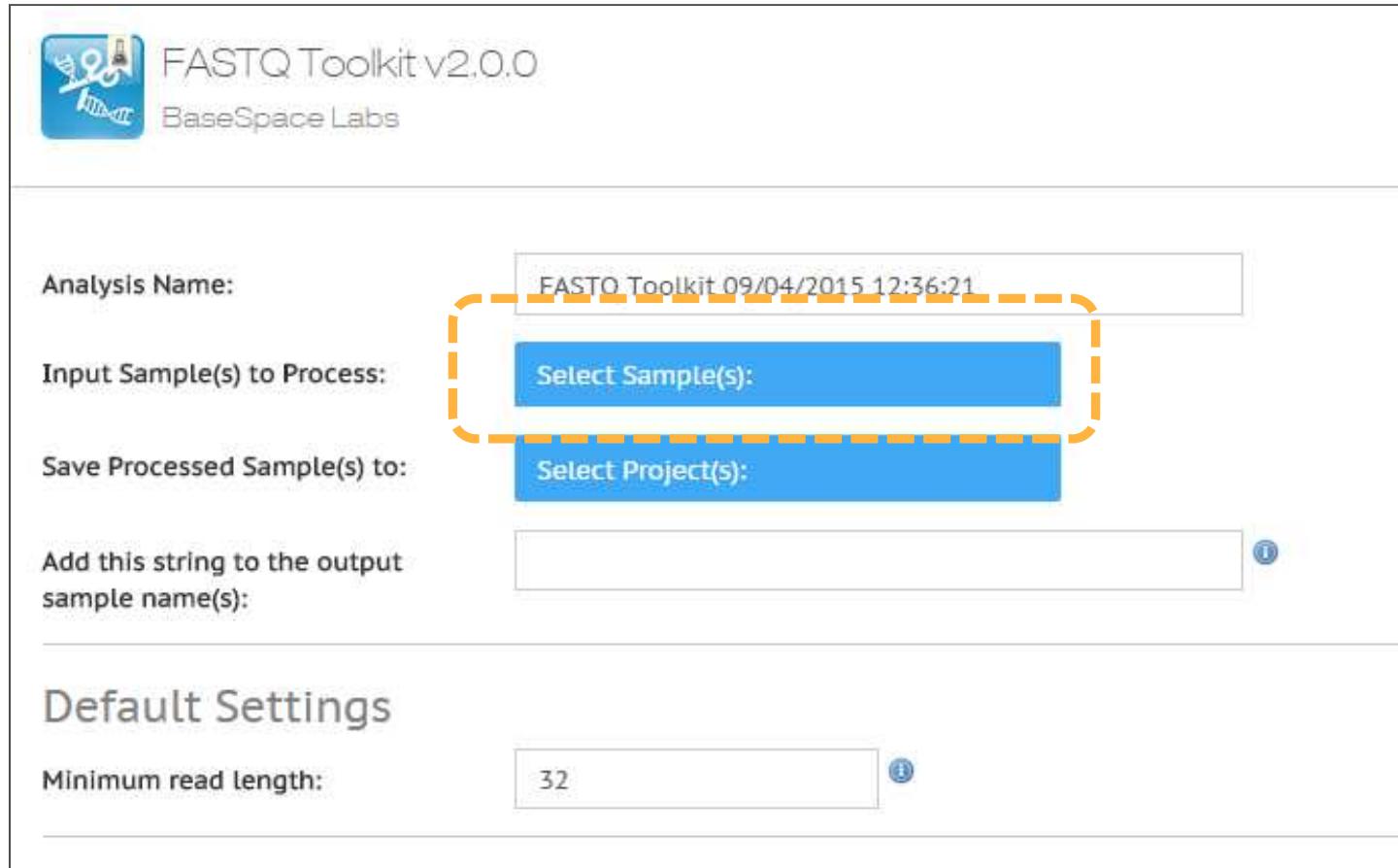
- Sample1\_S1\_L001\_R2\_001.fastq.gz
- Sample1\_S1\_L001\_R1\_001.fastq.gz

A 'mamma' logo is visible at the bottom right of the dialog.

# FASTQ のアップロード



# 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



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/misqreporter/misq-reporter-generatefastq-workflow-guide-15042322-b.pdf](http://support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/misqreporter/misq-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](http://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](http://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。