NGS解析をはじめよう
～よりよいデータ解析のためのFASTQファイルの前処理とデータ評価～

フィールドアプリケーション部
木村成子

2020年5月27日

For Research Use Only. Not for use in diagnostic procedures.

illumina

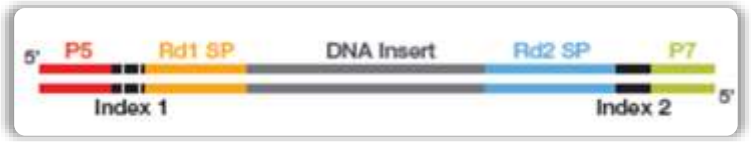
ウェビナーの概要

- **データ解析ワークフロー**（FASTQファイル作成までとその後の2次解析）
  - データ解析フローの概要
  - FASTQ ファイルフォーマット
- **FASTQ ファイル前処理の種類**
  - アダプタートリミング
  - クオリティトリミング
  - リードの結合
  - サブサンプリング
- **FASTQ ファイル前処理の具体例**
  - Fastqファイル作成と同時にアダプタートリミングする方法
  - BaseSpace Sequence Hub (BSSH) FASTQ Toolkit App を用いた前処理
- **FASTQ ファイルのクオリティ評価**
  - BaseSpace Sequence Hub (BSSH) FastQC App を用いたクオリティ評価

データ解析のワークフロー

illumina

データ解析フロー



- シーケンス画像取得 / シグナル抽出
- ベースコール（塩基の決定）

*.bcl ファイル

FASTQ ファイル

- 装置内ソフトウェア（Local Run Manager、MiSeq Reporter）
- BaseSpace Sequence Hub (BSSH)
- Bcl2fastq（Linux）

- 2次解析ワークフロー
- Third Party

BAMs, VCFs, Assembly

illumina

FASTQ ファイル

- **ファイル名**
  SampleName_SampleNumber_Lane_Read_FlowCellIndex.fastq.gz
  SampleName_S1_L001_R1_001.fastq.gz
  SampleName_S1_L001_R2_001.fastq.gz

- **FASTQ ファイルフォーマット**

```
@SIM:115:FCX:4:2106:6329:1045 ATGCTAGCC 1:N:0:ATCCGA
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;BB=><9=AAAAAAAAAA9B:<B<;<<<???B=
```

ヘッダー

塩基配列

クオリティスコア

- **ヘッダー名**

  @Instrument:RunID:FlowCellID:Lane:Tile:X:Y:UMI Read:Filter:0:IndexSequence or SampleNumber

2020/03/25
NGS解析をはじめよう ～基礎からわかるバイオインフォマティクス入門編～
テクニカルアプリケーションサイエンティスト　仁田原 翔太
https://jp.illumina.com/events/webinar/2020/webinar-0325-j.html

illumina

クオリティスコア（Qスコア）

- クオリティスコアは、各塩基のベースコールの確からしさを表す指標。データ評価を行う時に用いる。
  - Q10: 90 %の確率でそのベースコールは確からしい
  - Q20: 99 %の確率でそのベースコールは確からしい
  - Q30: 99.9 %の確率でそのベースコールは確からしい
  - Q40: 99.99 %の確率でそのベースコールは確からしい

Q-Score	Symbol	Q-Score	Symbol	Q-Score	Symbol
0	!	14	/	28	=
1	“	15	0	29	>
2	#	16	1	30	?
3	$	17	2	31	@
4	%	18	3	32	A
5	&	19	4	33	B
6	‘	20	5	34	C
7	(	21	6	35	D
8	)	22	7	36	E
9	*	23	8	37	F
10	+	24	9	38	G
11	,	25	:	39	H
12	-	26	;	40	I
13	.	27	<		

データ解析のワークフロー
～FASTQファイルが手に入ったら～

FASTQ ファイル

- 装置内ソフトウェア（Local Run Manager、MiSeq Reporter）
- BaseSpace Sequence Hub (BSSH)
- Bcl2fastq（Linux）

① FASTQファイルの特徴を知る！(データ評価）
→ クオリティスコア分布
リード長分布
→ FastQC App

② FASTQファイルの前処理
→ FASTQ Toolkit App

③ FASTQファイルの確認
→ FastQC App

BAMs, VCFs, Assembly

- 2次解析ワークフロー
- Third Party

FASTQファイル前処理の種類

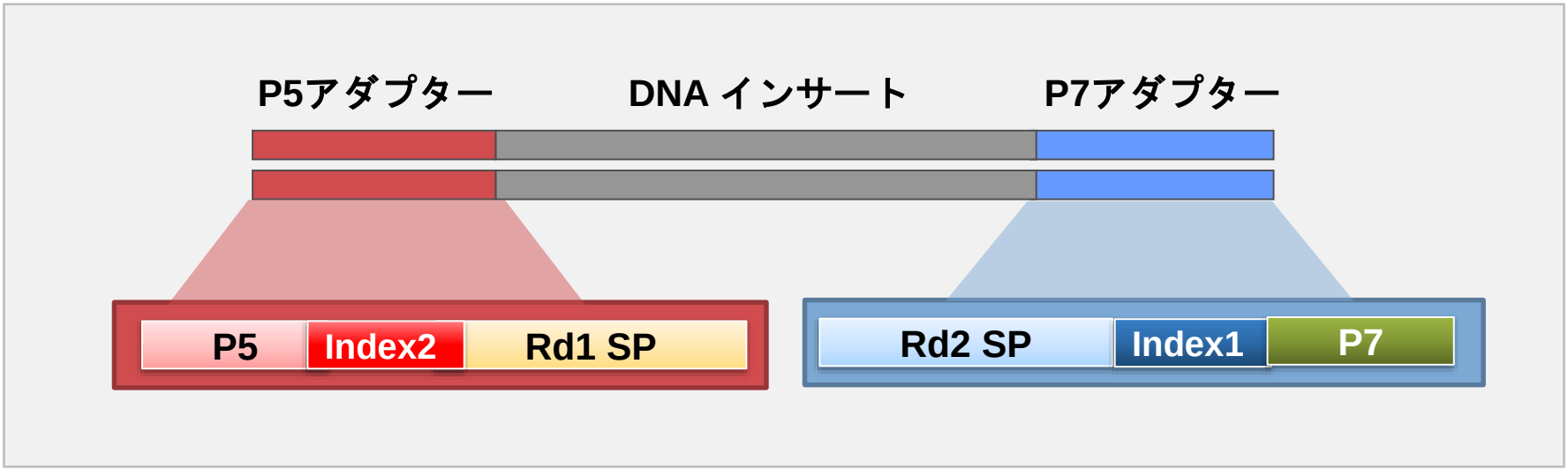
illumina

アダプタートリミング

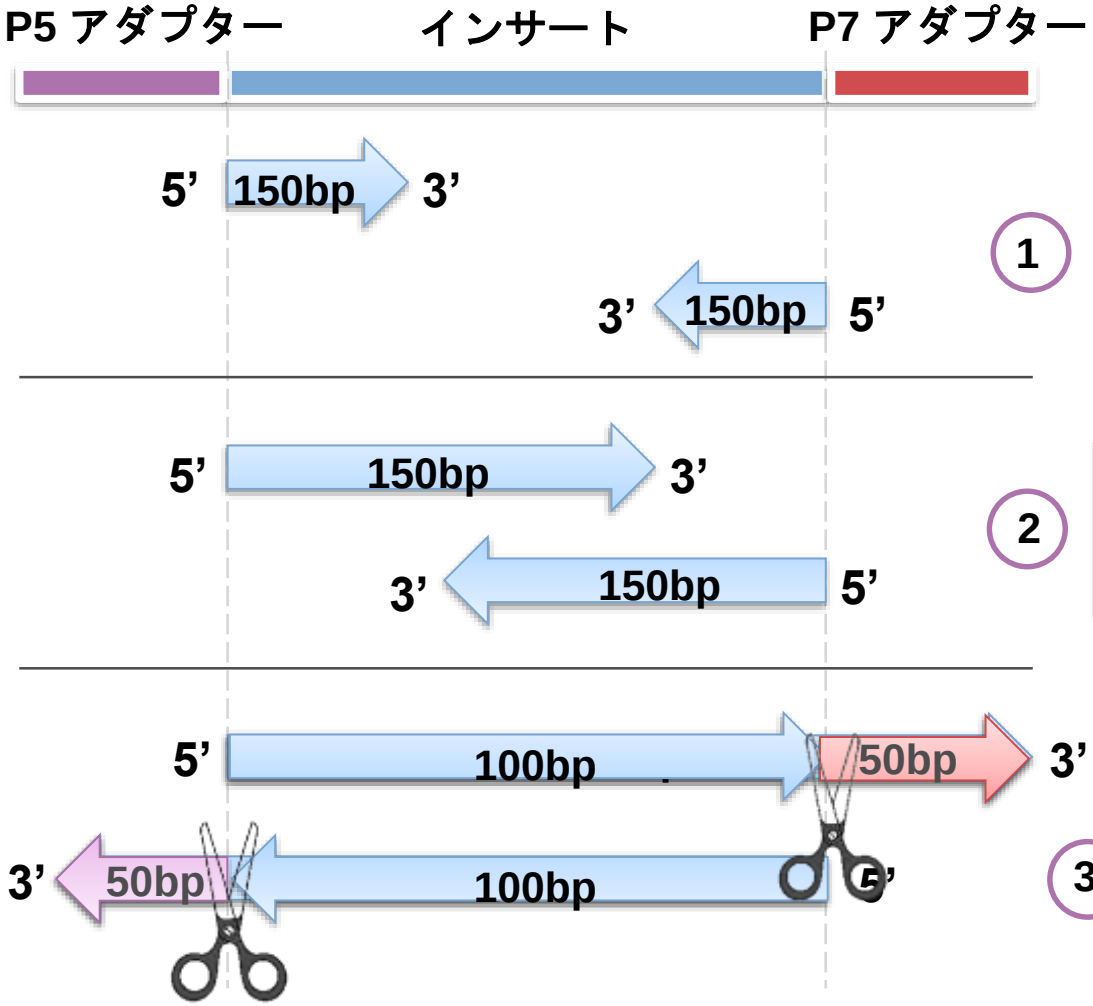
illumina

ライブラリ構造
ーアダプターとは？ー

- ライブラリ ＝　DNA インサート　＋　アダプター
- Read 1 シーケンスプライマー (Rd1 SP)
- Read 2 シーケンスプライマー (Rd2 SP)



アダプタートリミング
例：150 bp x 2



1
インサート長= 1000 bp
リード長 ＜ インサート長

2
インサート長 = 200 bp
リード長 ＜ インサート長
（部分的に、オーバーラップ）

3
インサート長 = 100 bp
リード長 ＞ インサート長

illumina

アダプタートリミング（ハードトリミング）

アダプター配列
マッチ> 90%
(デフォルト)

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


@M00000:71:000000000-D00LW:1:1101:16265:1658 1:N:0:1
ACTCTGCGTTGCGCTTCTGCTCGGCCTCCAGCTCACCCTCC
+
BCCCCFFCCBCCGGGGGGGGGGGGGGGHHHHHHHHHHHHHG
```

アダプター配列以降を完全に除去する

illumina

アダプタートリミング
（アダプターマスキング）

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

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

アダプター配列をNでマスキングし、クオリティスコアを２（#）に置き換える
（もし、後続のデータ解析で全て同じリード長の配列が必要な時に使用）

どうしてアダプタートリミングが必要?

1

Higher alignment %

BWA
(backtrace)

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

illumina

どうしてアダプタートリミングが必要?

2 Improved assemblies

Data: 2 x 250bp, *E.coli* ゲノムサイズ：4.64M (Nextera™ XT)

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

N50：アッセンブリした配列を長い順に並べて、長い方から順に配列の塩基数を足していった時に全体の長さの半分に達したときの配列の塩基数(bp)

Number of contig：アッセンブリしてつなぎ合わされた配列の数

クオリティトリミング

illumina

クオリティトリミング

- **平均クオリティスコアを基に、リード両末端のクオリティの低い塩基を取り除く**
- **どんな時にトリミングする?**
  - 基本的に、全てのアプリケーション
  - 特に3'末端のクオリティの低い塩基が、後続の解析にクリティカルに影響を与える時　（e.g. de novo アセンブリ, リード結合など）
- **どんな時にトリミングしない?**
  - リシーケンス解析： ほとんどのソフトウェア (i.e. BWA, Isaac) がクオリティスコアを考慮に入れてアライメントしているため。
- **イルミナツール**
  - BSSH FASTQ Toolkit App

クオリティトリミングの例

QualityScoreTrim,20

```
@M00000:72:000000000-D00LW:1:1101:22420:18334 1:N:0:1
CACCAAGGGCCTGGGGTGTCAATGGCGGGGCTTGTGACTGCACAAAAGGGGCCTCCCGCAGGGGCTCCCGCC
+
BBBBBBFBBBBBGGGGEEFGGGHHHHGGG00>10B355@BB3@3BG1?E1///1/11/////////////////
```

36

```
@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
24	9
30	?
31	@
32	A
33	B
36	E

クオリティトリミングの例

QualityScore:20, WindowLength:4

20 x 4 = 80 >75

```
@M00000:72:000000000-D00LW:1:1101:22420:18334 1:N:0:1
CACCAAGGGCCTGGGGTGTCAATGGCGGGGCTTGTGACTGCACAAAAGGGGCCTCCCGCAGGGGCTCCCGCC
+
BBBBBBFBBBBBGGGGEEFGGGHHHHGGG00>10B355@BB3@3BG1?E1///1/11/////////?////
```

Window

97 75

```
@M00000:72:000000000-D00LW:1:1101:22420:18334 1:N:0:1
CACCAAGGGCCTGGGGTGTCAATGGCGGGGC
+
BBBBBBFBBBBBGGGGEEFGGGHHHHGGG00
```

Q	ASC
13	.
14	/
15	0
16	1
18	3
20	5
22	7
24	9
30	?
31	@
32	A
33	B

illumina

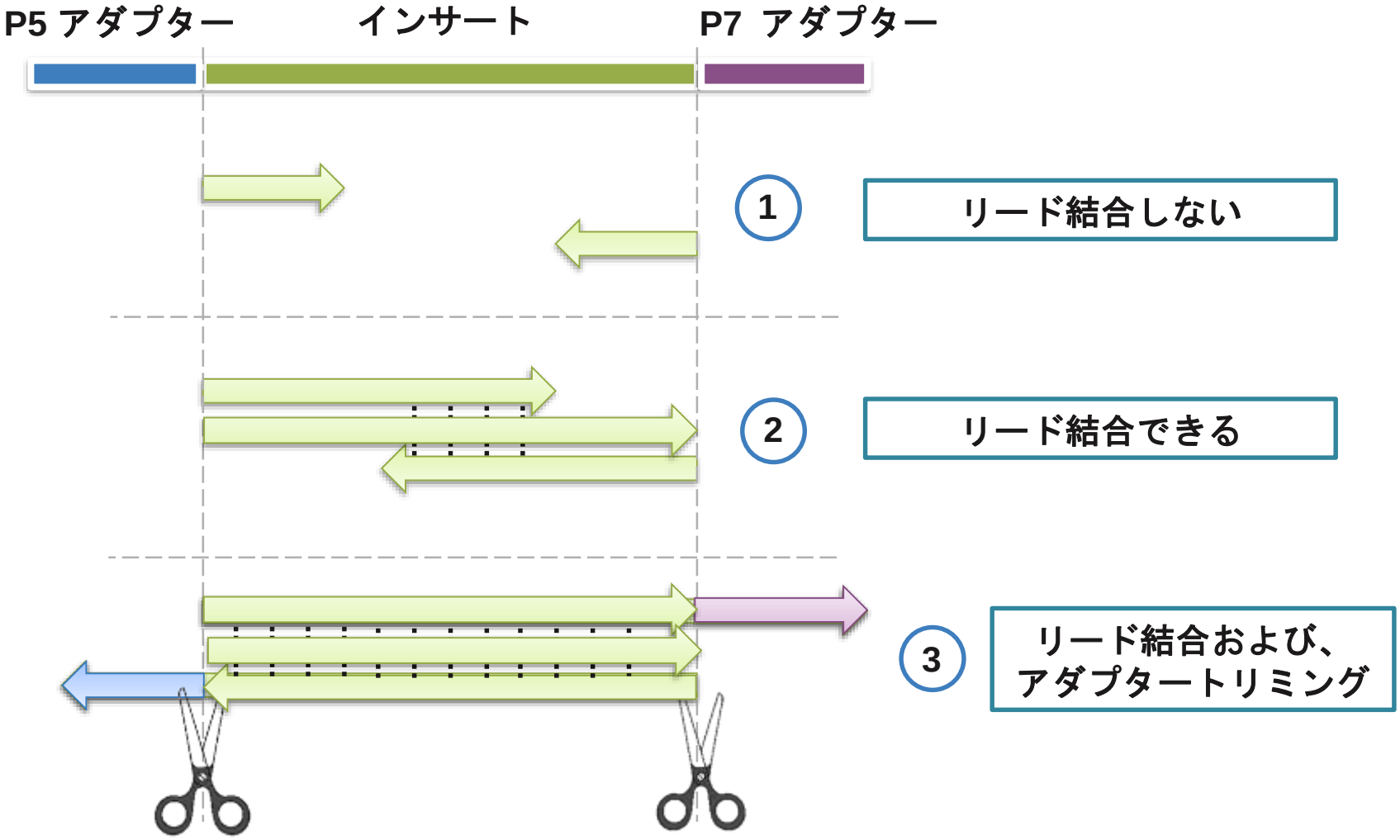
リード結合

illumina

リード結合

- **ペアードエンドのR1とR2をつなぎ合わせ、シングルリードにすること**
- **どんな時にリード結合が必要?**
  - より長いリードを得ることが重要な時
  - シングルリードしか受け付けないソフトウェアを使用する時
  - 大部分のリードがオーバーラップしている時
  - オーバーラップ領域にindelが検出される時
- **どんな時にリード結合が必要ではない?**
  - 多くのリードにオーバーラップ領域がないとき
  - オーバーラップ領域に繰り返し配列があるとき
  - 後続の解析で、R1、R2ファイルを別々にインプットする必要があるとき
- **一般的に用いられるツール**
  - USEARCH, FLASH

リード結合



illumina

サブサンプリング

illumina

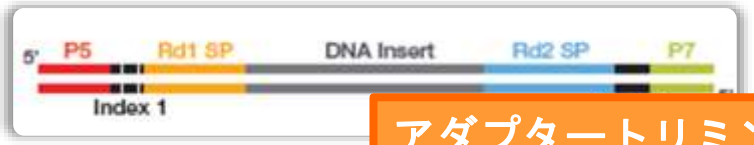
サブサンプリング

- **シーケンスリード数が多すぎるときに、ランダムに一部のリードのみを抜き出してサブサンプルセットを作成すること**
- **どうしてサブサンプルが必要?**
  - トラブルシューティングのための解析をより短い時間で行うため
  - パソコンの性能が十分ではなく、シーケンスデータが大きすぎて、データ解析が途中でタイムアウトする場合
  - イルミナBaseSpace Sequence Hub のいくつかのアプリには、インプット量に制限があるため
  - 解析に適当なリード数の検討を行う場合
- **イルミナツール**
  - BSSH FASTQ Toolkit App

FASTQファイルの前処理の方法
―具体例―

illumina

データ解析フロー



アダプタートリミング
クオリティトリミング
リード結合
サブサンプリング

・シーケンス
　画像取得 /
　シグナル抽出
・ベースコール
　（塩基の決定）

*.bcl ファイル

アダプタートリミング

FASTQ ファイル

- 装置内ソフトウェア
  （Local Run Manager、MiSeq Reporter）
- BaseSpace Sequence Hub (BSSH)
- Bcl2fastq（Linux）

- 2次解析
  ワークフロー
- Third Party

BAMs, VCFs, Assembly

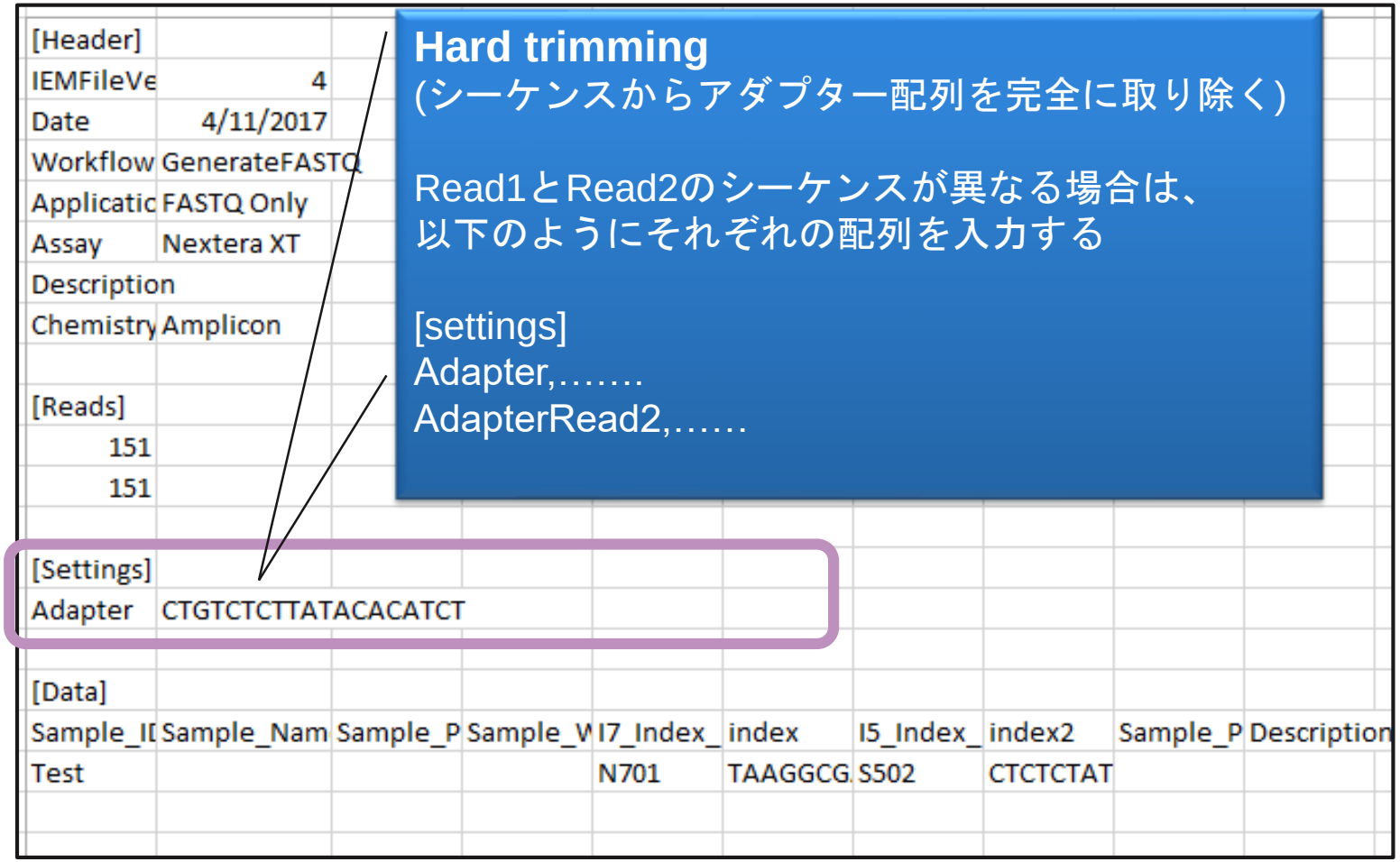
illumina

FASTQファイル作成時にアダプタートリミングする方法

illumina

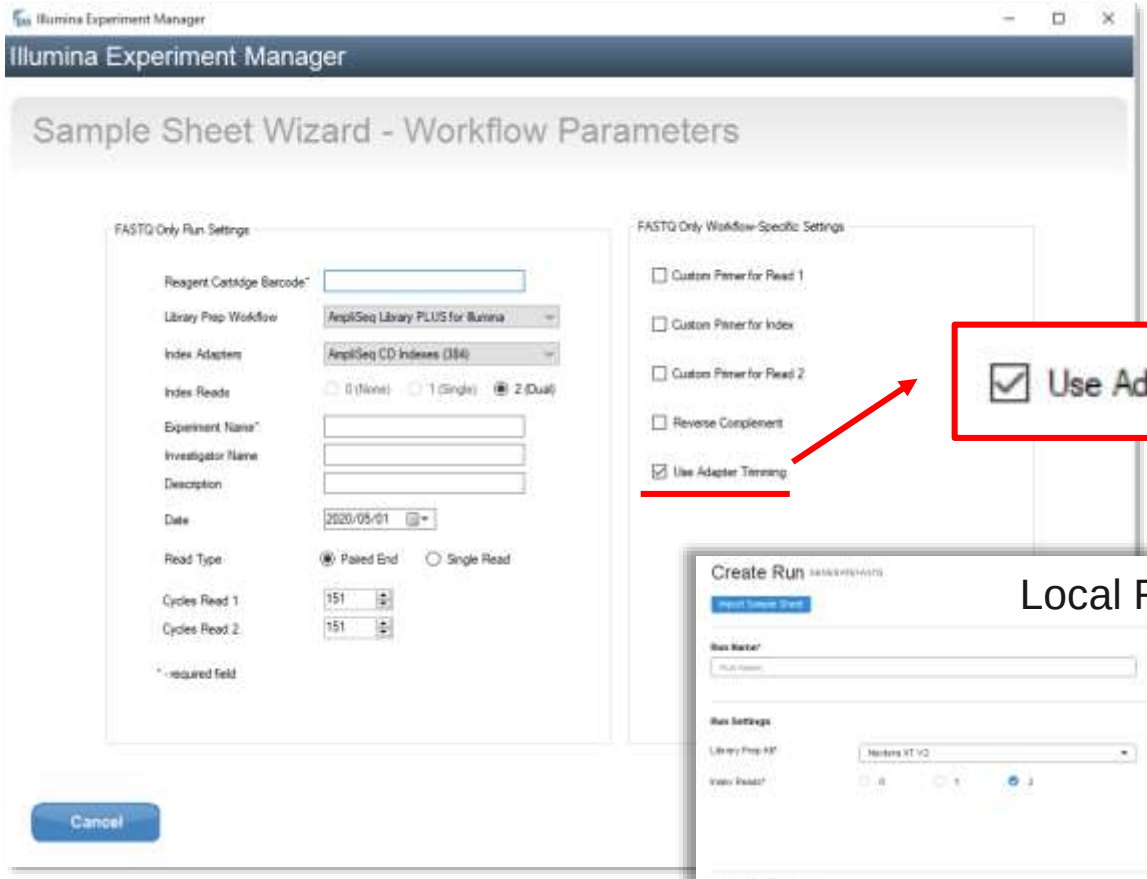
アダプタートリミング

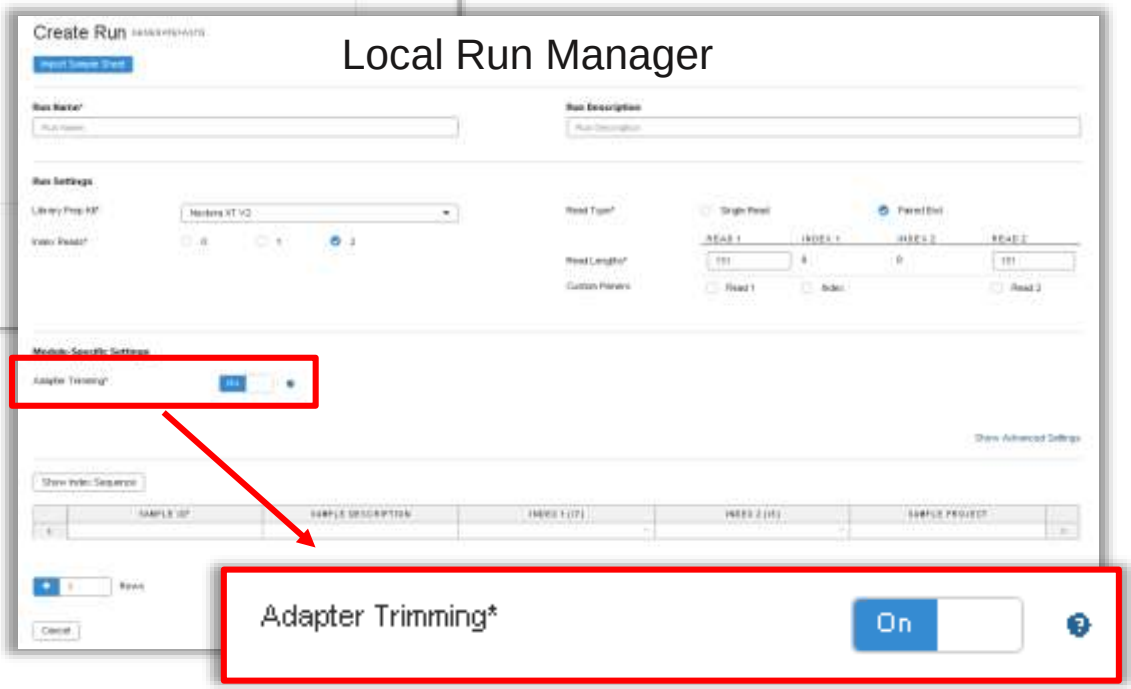
MiSeq™ Reporter, Local Run Manager, BaseSpace™ Fastq Generation and Bcl2fastq



https://support.Illumina.com/bulletins/2016/12/what-sequences-do-i-use-for-adapter-trimming.html

illumina





2018/07/25
サンプルシート作成ツール：
Illumina Experimental Manager（IEM）の使用方法 -
最新バージョンIEMv1.15のご紹介 –
【イルミナiSchool 初級】
フィールド アプリケーション サイエンティスト
山重りえ・上利佳弘
https://jp.illumina.com/events/webinar/2018/webinar-180725-j.html

illumina

アダプターマスキング

Bcl2fastq

[Header]									
IEMFileVersio	4								
Date	4/11/2017								
Workflow	GenerateFASTQ								
Application	FASTQ Only								
Assay	Nextera XT								
Description									
Chemistry	Amplicon								
[Reads]									
151									
151									
[Settings]									
MaskAdapter	CTGTCTCTTATACACATCT								
[Data]									
Sample_ID	Sample_Nam	Sample_P	Sample_W	I7_Index_	index	I5_Index_	index2	Sample_P	Description
Test				N701	TAAGGCG	S502	CTCTCTAT		

N-masking
(アダプター配列を“N” 、それらのQスコアを“#”に置き換える)

Read1とRead2のシーケンスが異なる場合は、
以下のようにそれぞれの配列を入力する

[settings]
MaskAdapter,…..
MaskAdapterRead2,…..

BaseSpace Sequence Hub FASTQ Toolkit Appを用いたFASTQファイルの前処理

illumina

BaseSpace Sequence Hub

- BaseSpace Sequence Hub (BSSH) は、イルミナ社が提供するクラウドサービスで、イルミナ次世代シーケンサーのデータ管理及び解析にご利用いただけます（有償）。
- 90種類以上のアプリをご提供。解析結果は「project」に保存。
- アプリには、Core App、 BaseSpace Lab App、Third Party Appの3種類。

Core App

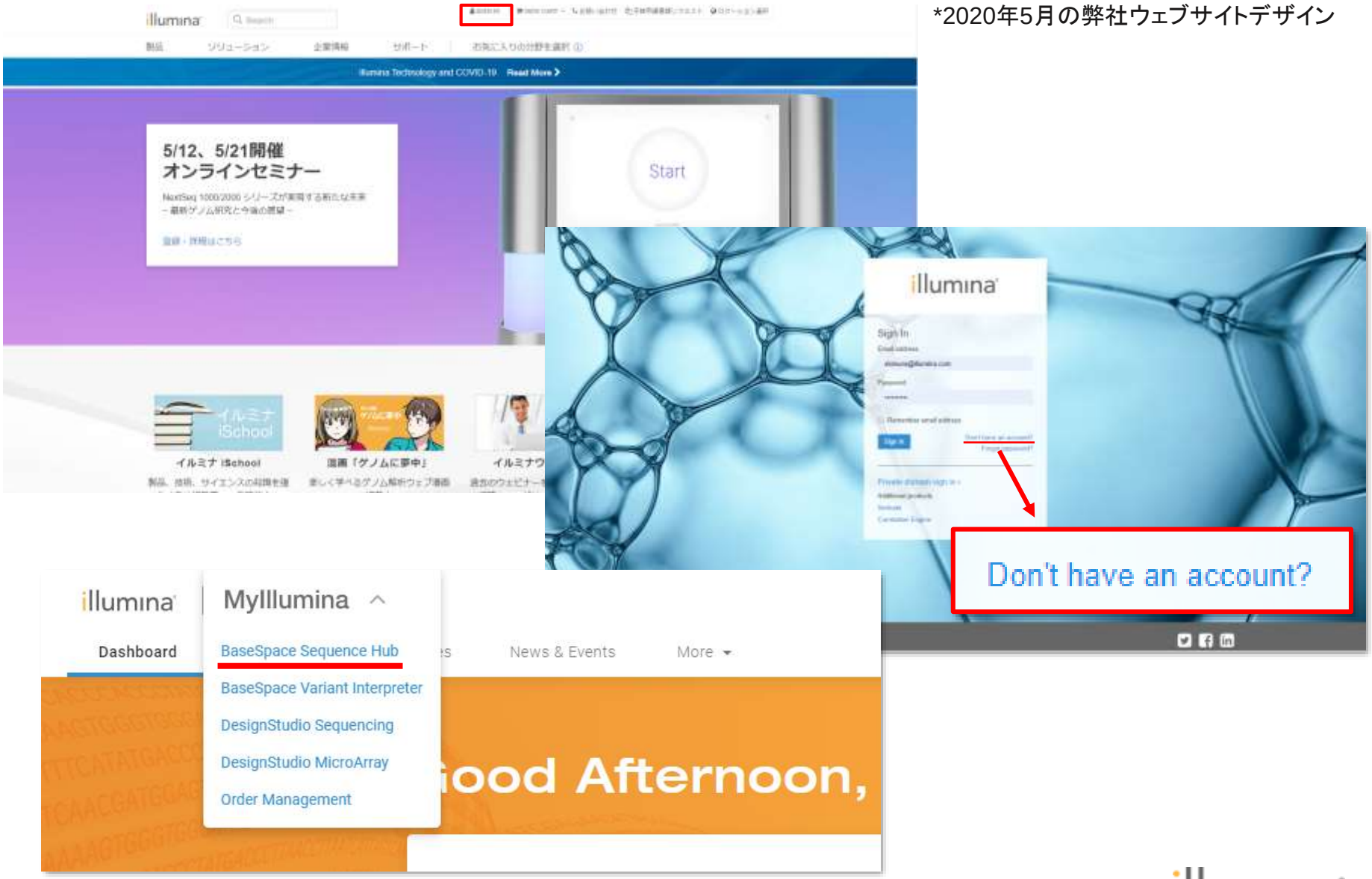
BaseSpace Lab App

Third party App

2017/11/29
BaseSpace Sequence Hubを使用したデータ解析の基本【イルミナiSchool 初級】
テクニカル アプリケーション サイエンティスト　渡邊　大
https://jp.illumina.com/events/webinar/2017/webinar_171129_j.html#webinarFormLocation

BaseSpace Sequence Hubのアカウント取得

*2020年5月の弊社ウェブサイトデザイン

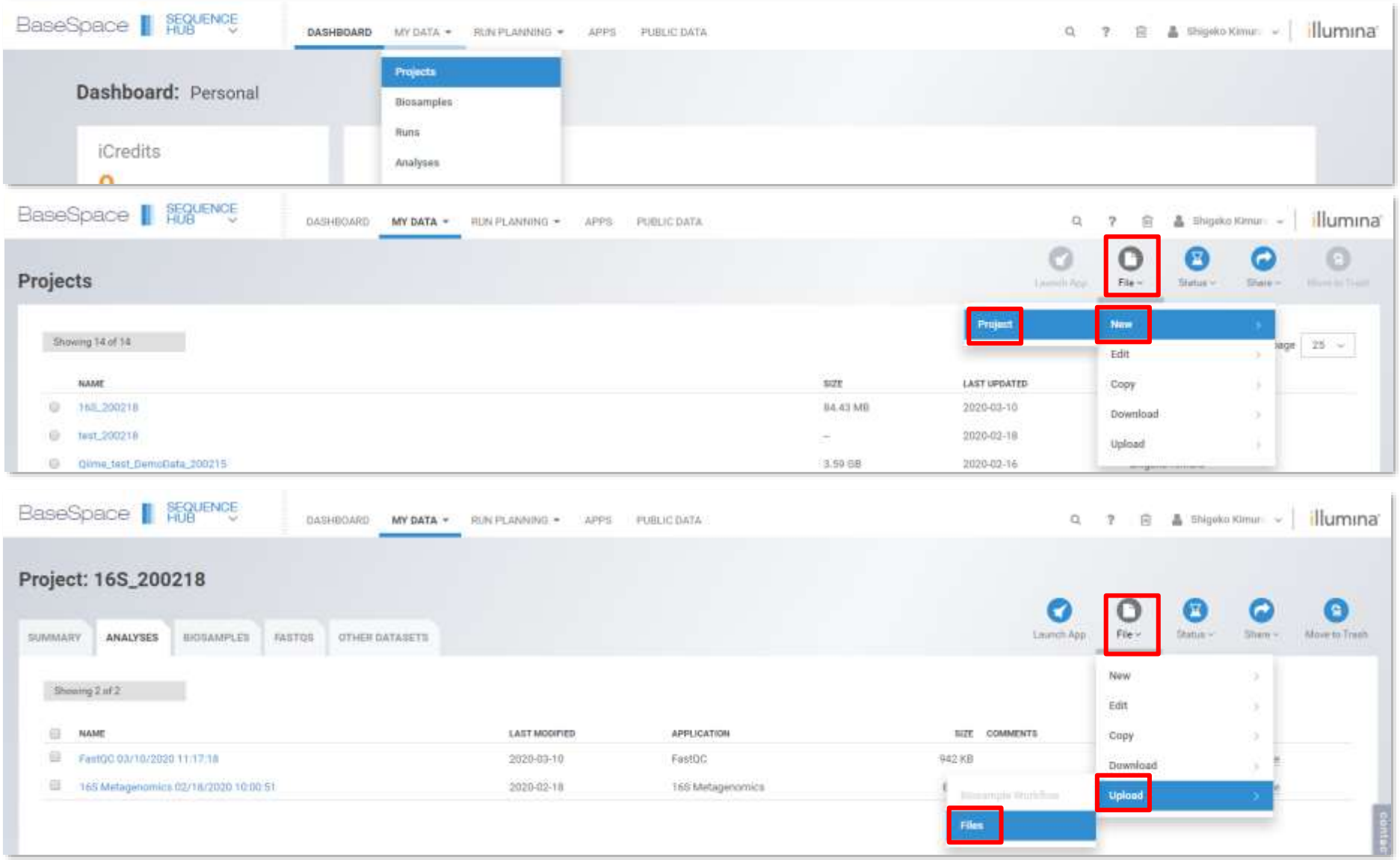


illumina

BaseSpace Sequence Hub App (>90種類)

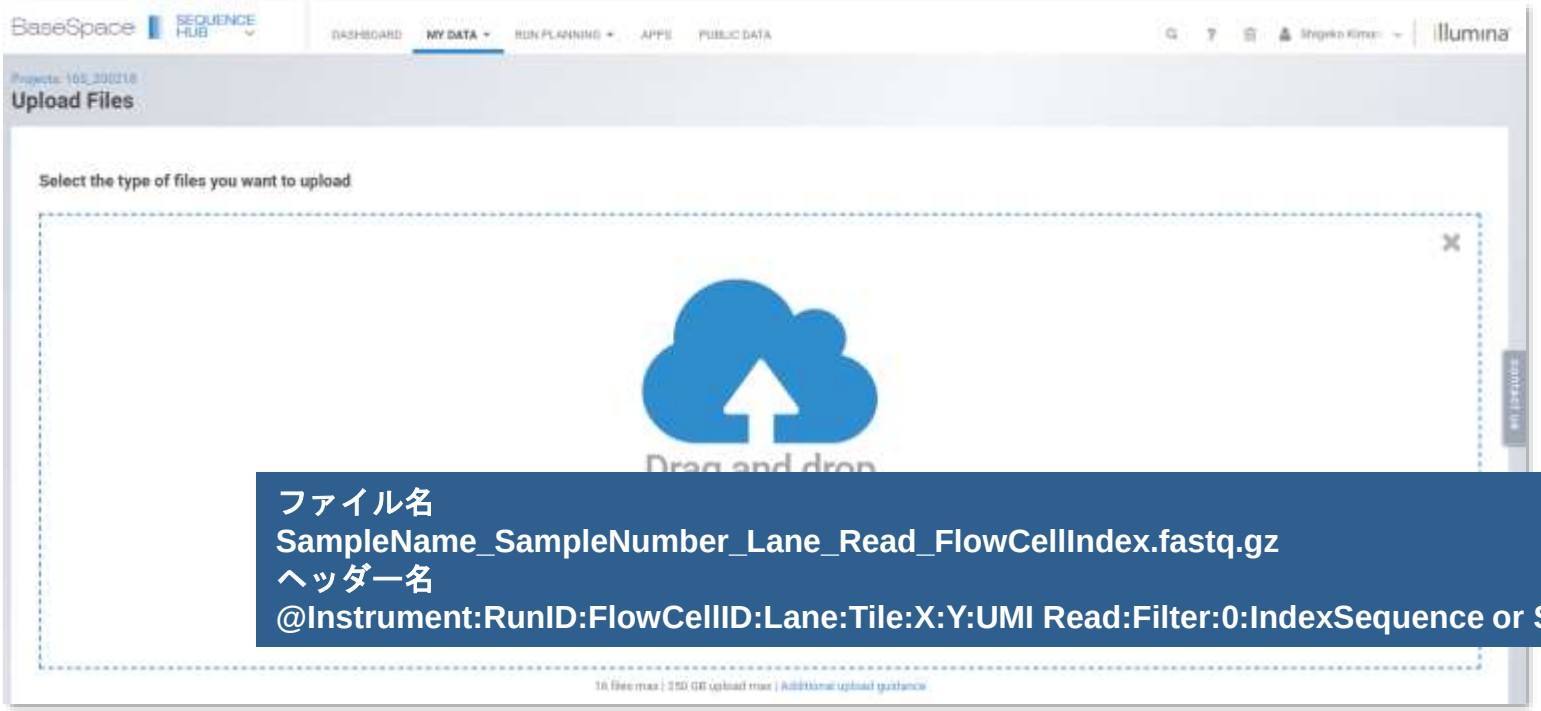
illumina

FASTQファイルのアップロード法



illumina

FASTQファイルのアップロード法

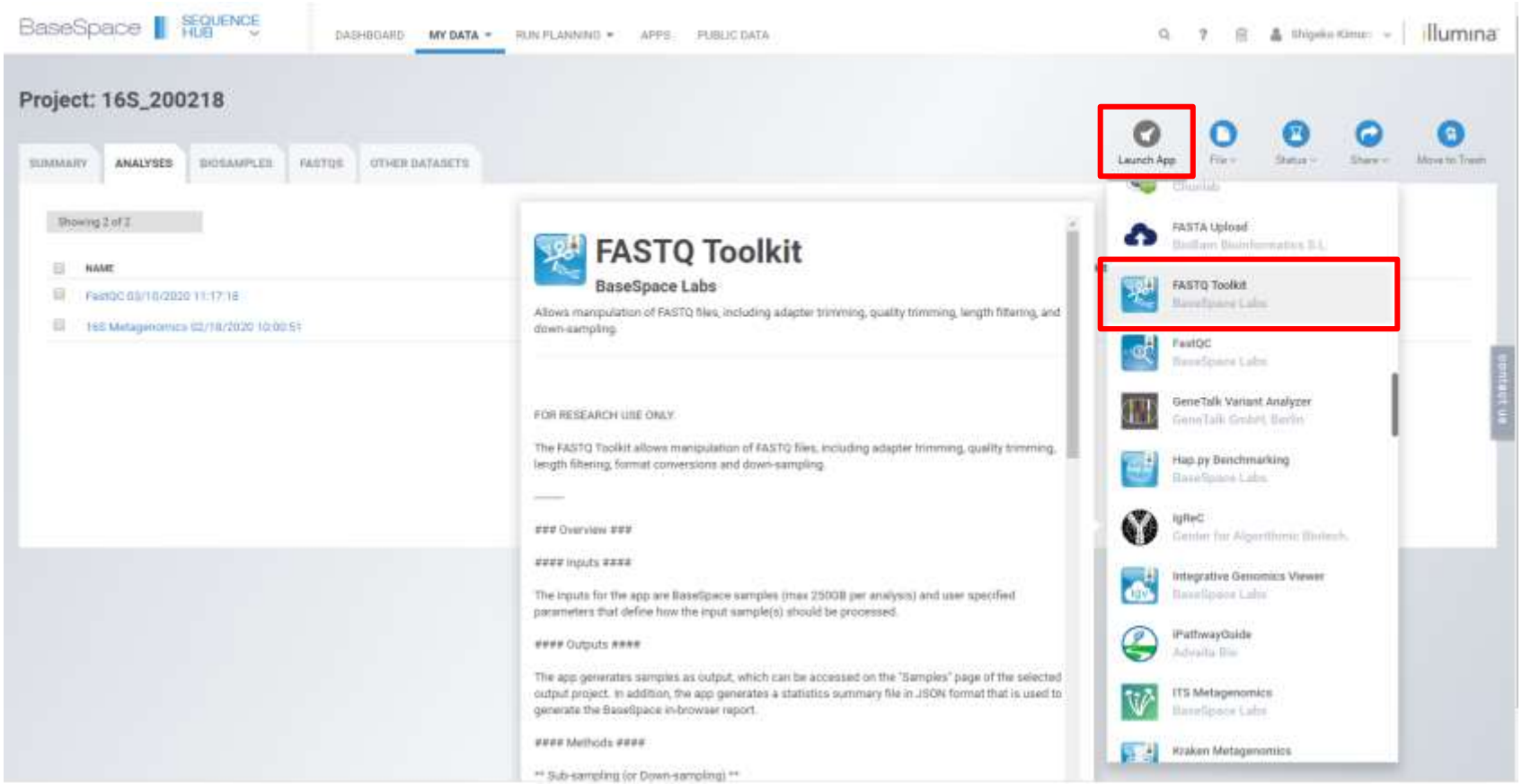


ファイル名
SampleName_SampleNumber_Lane_Read_FlowCellIndex.fastq.gz
ヘッダー名
@Instrument:RunID:FlowCellID:Lane:Tile:X:Y:UMI Read:Filter:0:IndexSequence or SampleNumber

illumina

FASTQ Toolkit App

- App→ “FASTQ Toolkit” Appアイコンを選択
- Project内、Launch app→“FASTQ Toolkit” Appを選択



illumina

FASTQ Toolkit App

FASTQ Toolkit v2.2.0
BaseSpace Labs

Configuration

Analysis Name

FASTQ Toolkit 04/29/2020 4:45:20

Input Sample(s) to Process

Select Sample(s):

ATCCiSeq1002x150Rep7

Save Processed Sample(s) to

Select Project

test

Add a suffix to the output sample name(s)

trim

結果ファイル名は、インプットファイル名の後ろに、
この欄に記入した文字が付加されたものです。
空欄の場合、結果ファイル名がインプットファイル名と
同一になりますのでご注意ください！

Please Note

Output sample names are required to be unique. Above, you can add a suffix to the input sample names or select a project different from the input sample project.

Default Settings

Minimum read length

32

illumina

FASTQ Toolkit App

- 適当なOptionを選択

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.

+ **Click to expand/hide sub-sampling settings**

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.

+ **Click to expand/hide adapter trimming settings**

Base Trimming

Bases can be trimmed from either the 5'- or 3'-end by specifying a number of bases for each end. Alternatively, bases can be trimmed from the 3'-end by specifying a maximum read length after trimming (e.g. to trim 150bp reads to 100bp reads).

+ **Click to expand/hide base trimming settings**

Quality Trimming

Use these parameters to trim low quality bases. Please note that most aligners (including BWA and Isaac) perform quality trimming internally during alignment. Quality trimming is more appropriate for workflows such as assembly, metagenomics or Methyl-Seq.

+ **Click to expand/hide quality trimming settings**

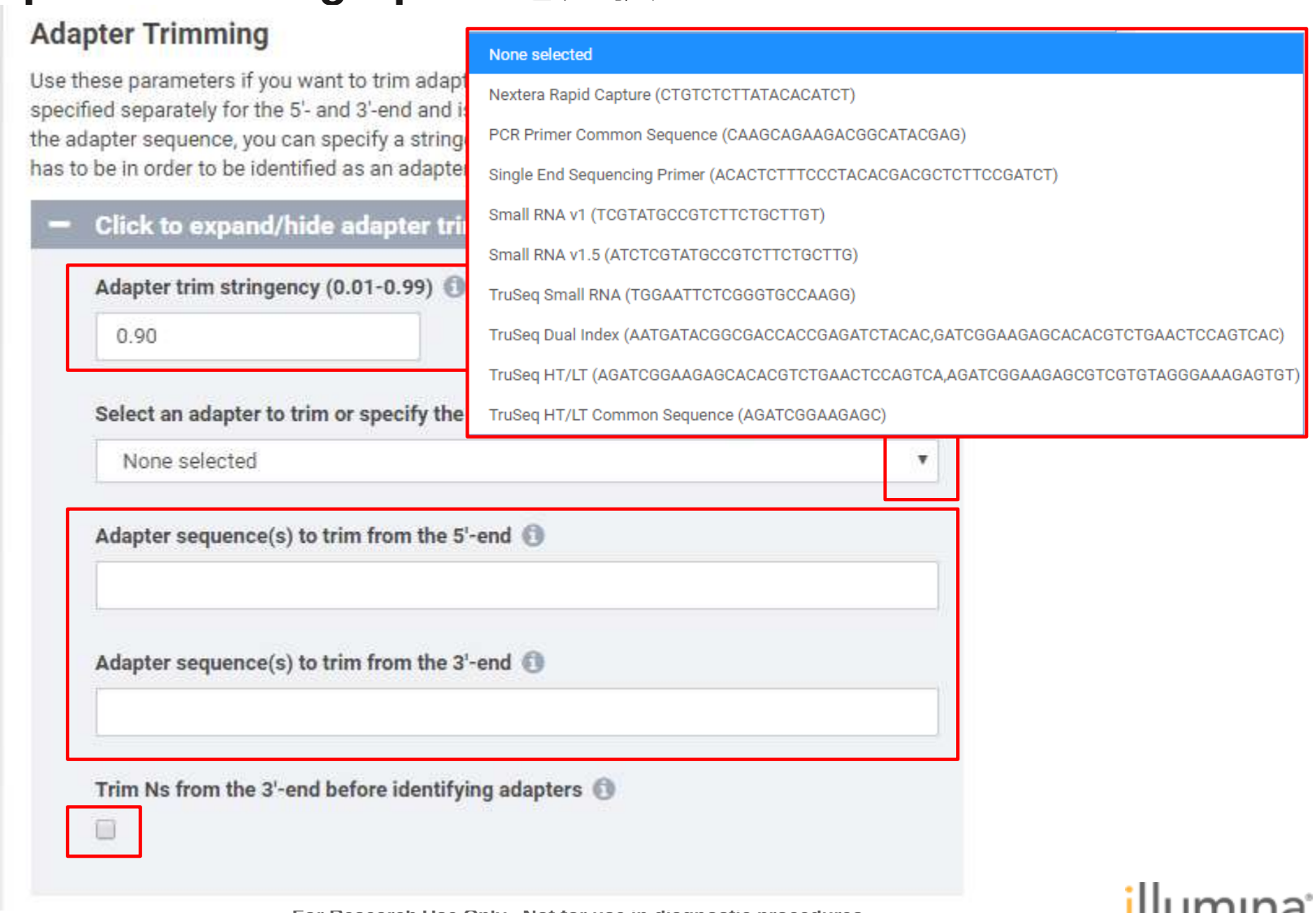
Poly-A/T Trimming

Poly-A/T tails are considered repeats of As or Ts at the read ends. The minimum length of a tail can be specified below. A small number of tails can occur even after trimming poly-A/T tails. For example, a sequence that ends with AAAAATTTTT and that has been trimmed for the poly-T will still contain the poly-A. Trimming poly-A/T tails can reduce the number of false positives during database searches, as long tails tend to align well to sequences with low complexity or sequences with tails (e.g. viral sequences) in the database. It also improves assemblies of (meta-)transcriptomes.

illumina

FASTQ Toolkit App

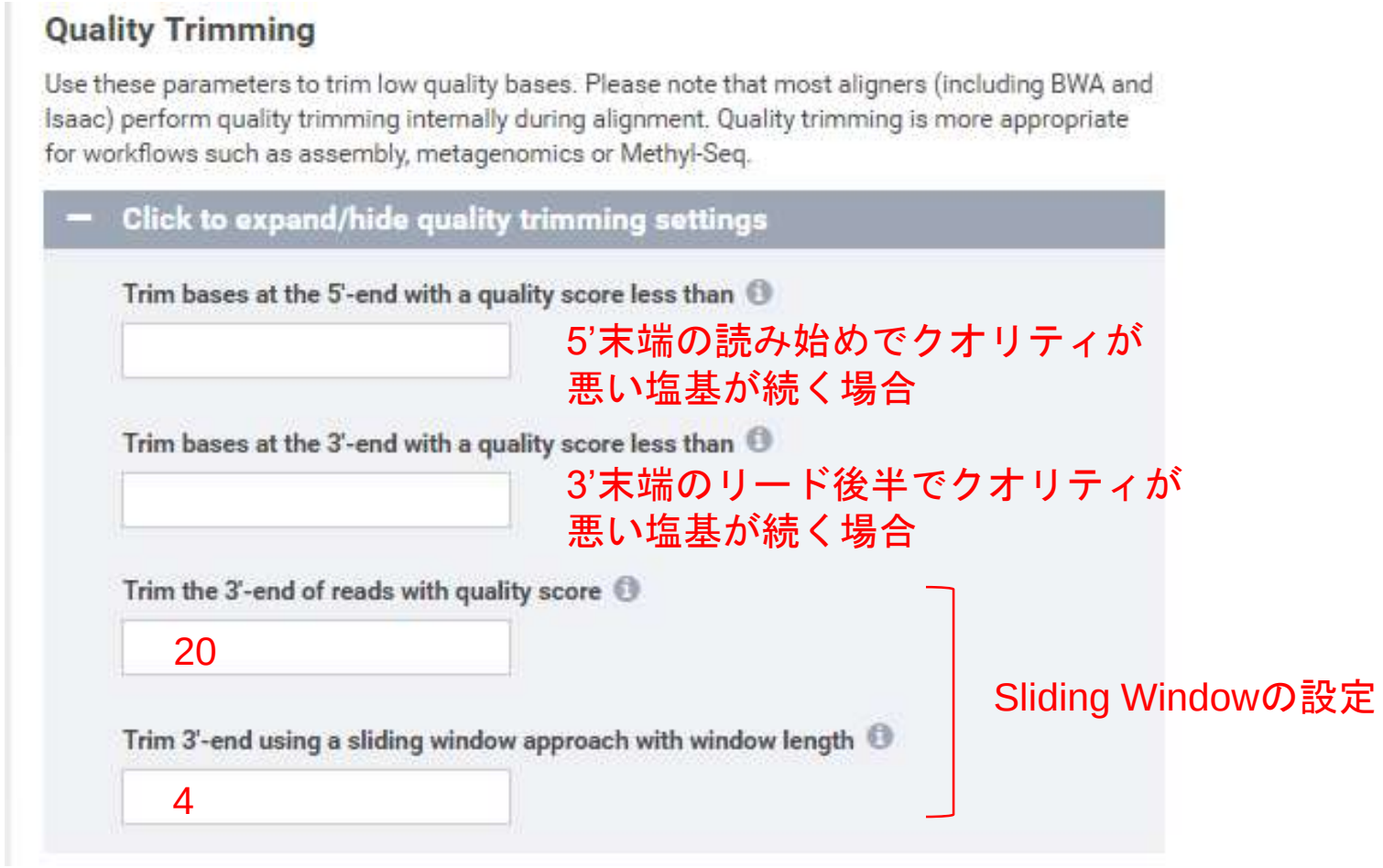
- Adapter Trimming optionを選択



illumina

FASTQ Toolkit App

- **Quality Trimming optionを選択**



illumina

FASTQ Toolkit App

- **Sub-Sampling optionを選択**

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.

Click to expand/hide sub-sampling settings

Maximum number of FASTQ entries to keep
絶対値

Maximum percentage of FASTQ entries to keep
割合

残したい最大リード数の設定

Maximum number of bases to keep
絶対値

Maximum percentage of bases to keep
割合

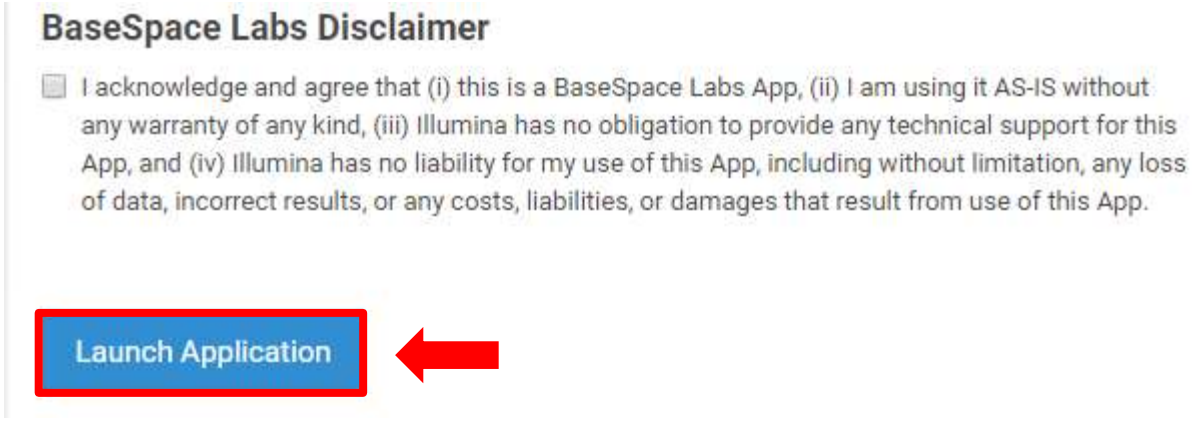
残したい最大塩基数の設定

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

illumina

FASTQ Toolkit App

- 条件設定が終わったら、規約に同意し、Launch Applicationを選択



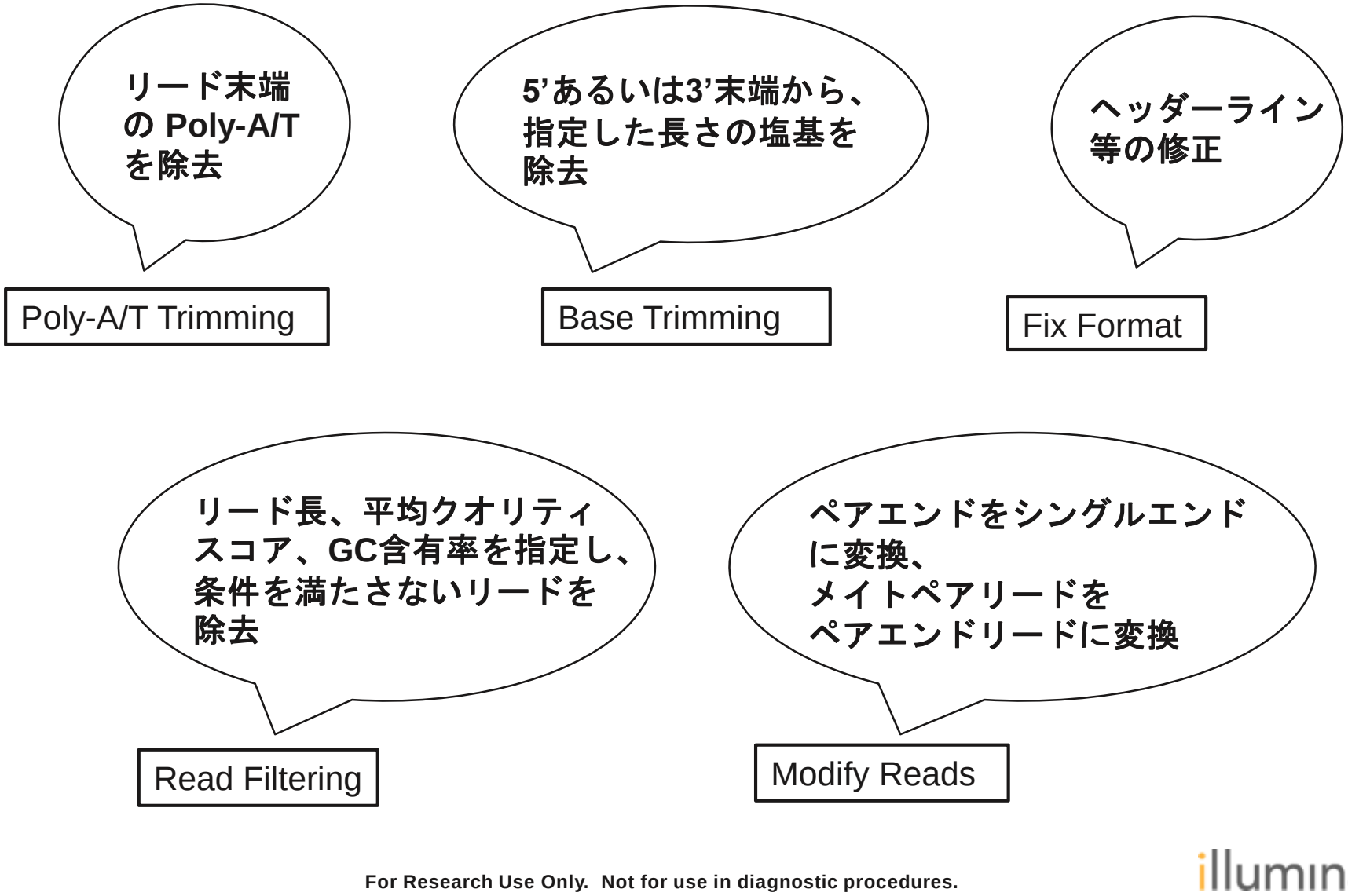
FASTQ Toolkit output- Results

original

Trimmed

illumina

FASTQ Toolkit その他のオプション



illumina

データ解析のワークフロー
～FASTQファイルが手に入ったら～

FASTQ ファイル

- 装置内ソフトウェア（Local Run Manager、MiSeq Reporter）
- BaseSpace Sequence Hub (BSSH)
- Bcl2fastq（Linux）

① FASTQファイルの特徴を知る！(データ評価）
→ クオリティスコア分布
リード長分布
→ FastQC App

② FASTQファイルの前処理
→ FASTQ Toolkit App

③ FASTQファイルの確認
→ FastQC App

BAMs, VCFs, Assembly

- 2次解析ワークフロー
- Third Party

illumina

シーケンスデータのクオリティ評価

illumina

私のシーケンスデータのクオリティはいい?

BSSH FastQC App

- 配列データのクオリティ評価を行うためのApp
  (サブサンプリングしたデータを使用すればより早く行える)
- 非常に様々なグラフが結果として出される。アプリケーションによって、チェックしなければならない項目が異なるので、いずれかの結果が”Failure“だからと言って、クオリティが悪いということではない。
- ライブラリの特徴を知ることができる
  - クオリティスコアの分布
  - リード長の分布（アダプタートリミングされている？）

過去のデータと比較

illumina

FastQC　入力画面

FastQC v1.0.0
BaseSpace Labs

Configuration

Output

App Session Name

FastQC 05/01/2020 11:19:33

Output Project

Select Project

Input

Input Sample

Select Sample(s):

Options

Kmer Size

5

Use Contaminant Filter

BaseSpace Labs Disclaimer

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.

Launch Application

illumina

FastQC: 基本情報

- 解析の基本情報を表で確認できる

  ファイル名、全リード数、リード長、GC含量

Basic Statistics

Measure	Value
Filename	ATCCiSeq1002x150Rep7_S7_L001_R2_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	314069
Filtered Sequences	0
Sequence length	151
%GC	52

illumina

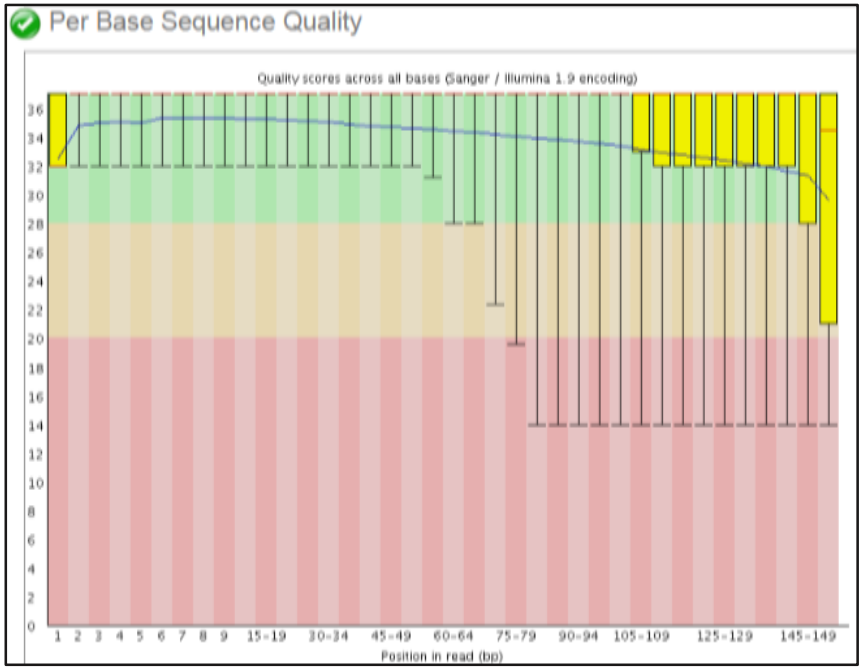
FastQC:
クオリティスコアの分布

各リードの平均クオリティスコアの分布

リードの平均クオリティスコア

平均クオリティスコアのピークはQ35
ほとんどがQ30以上
→クオリティの良いデータが得られている

クオリティスコアの分布



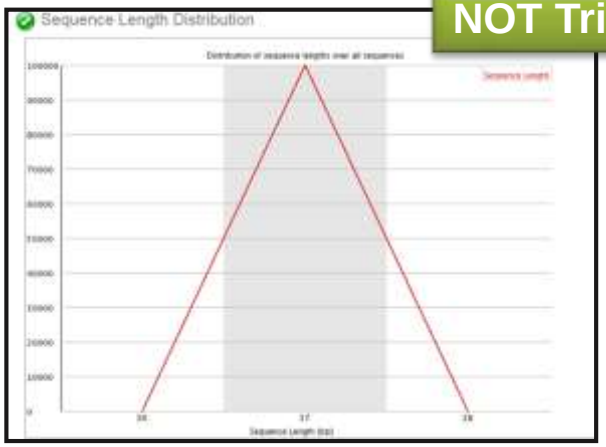
リード中の塩基の位置(bp)

クオリティスコアの分布
3’末端のクオリティスコアが落ちている
→クオリティトリミングを要検討

illumina

FastQC: リード長の分布

NOT Trimmed



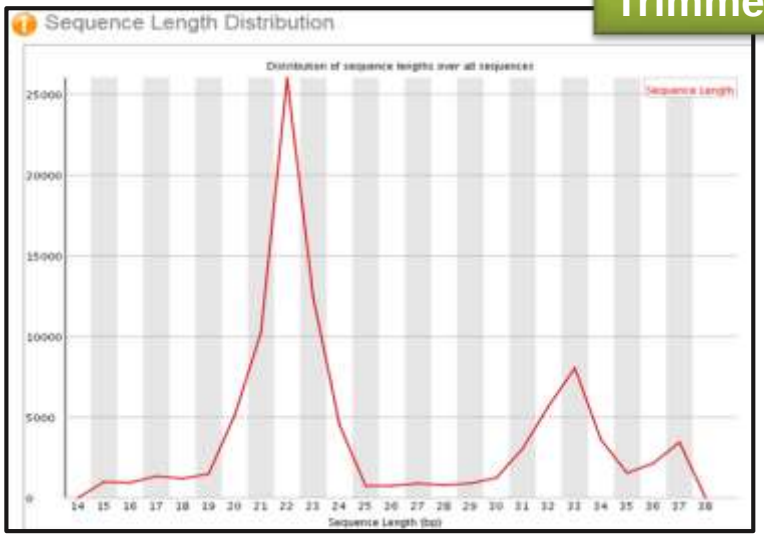
リード数

リード長 (bp)

- トリミングされていなければ、
  １つの大きなピークが見られる

例 1

Trimmed



リード数

リード長 (bp)

例 2

Trimmed

リード数

リード長 (bp)

参考資料

- FASTQ files explained
- What sequences do I use for adapter trimming?
- Adapter trimming: Why are adapter sequences trimmed from only the 3' ends of reads?
- FASTQC detailed documentation
- 【ウェビナー】NGS解析をはじめよう ～基礎からわかるバイオインフォマティクス入門編～
- 【ウェビナー】サンプルシート作成ツール：Illumina Experimental Manager（IEM）の使用方法 - 最新バージョンIEMv1.15のご紹介 –
- 【ウェビナー】 BaseSpace Sequence Hubを使用した　データ解析の基本

ご清聴ありがとうございました

illumina