Advances in Genome Science.

テクニカルサポートウェビナー 2013/10/11

「CASAVA でつくる FASTQ」

Eri Kibukawa

Bioinformatics Support Scientist

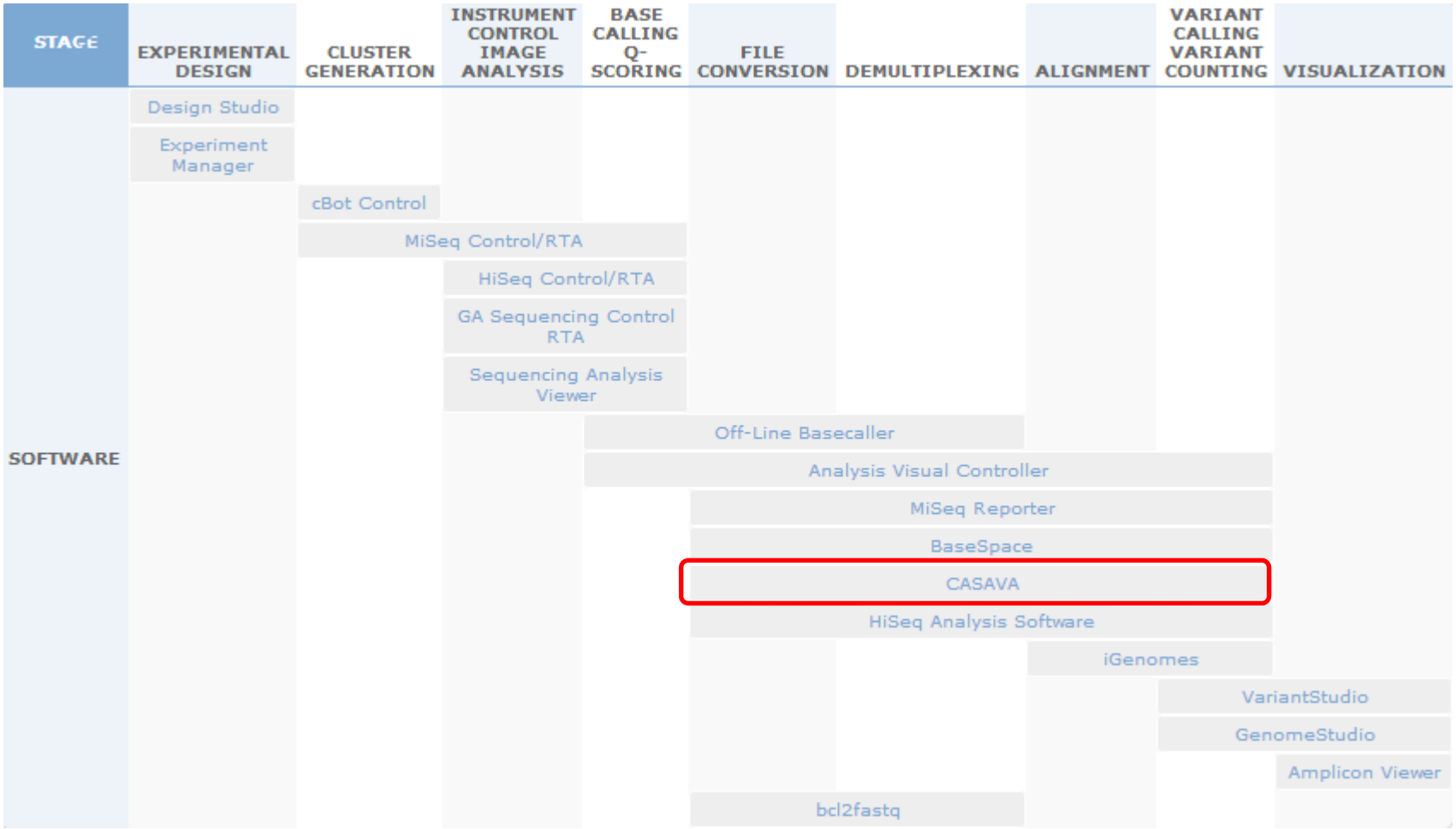
Illumina APAC

illumina®

本日お話しする内容

- HiSeq/MiSeq データをFASTQ変換する方法
- シーケンサーからのデータ転送
- ランフォルダ構造と主要ファイル
- 処理体系とコマンド
- configureBclToFastq.pl コマンドの使い方
- multiplexing のラン、インデクス配列
- CASAVA のサンプルシート
- コマンドライン デモ
- デマルチプレックス時のMSRとCASAVAの主な相違点
- HiSeq 1500/2500の圧縮機能を利用されている場合 ：bcl2fastq v1.8.4
- トラブルシュート時にお送り頂きたいファイル
- ドキュメントリソース

実験ステージ と イルミナシーケンシングソフトウェア



http://support.illumina.com/sequencing/sequencing_software.ilmn

illumina®

HiSeqとMiSeq データをFASTQ変換する方法の選択肢

選択肢	実行環境
CASAVA	ローカルLinuxで実行
BaseSpace	クラウド上のHAS (Hiseq Analysis Software) による自動処理 HASはローカルでも構築可能

選択肢	実行環境
BaseSpace	クラウド上のMSR(MiSeq Reporter) による自動処理
MiSeq Reporter	MiSeq上で自動処理 あるいはWindowsPCでボタン操作により実行
(CASAVA)	ローカルLinuxで実行

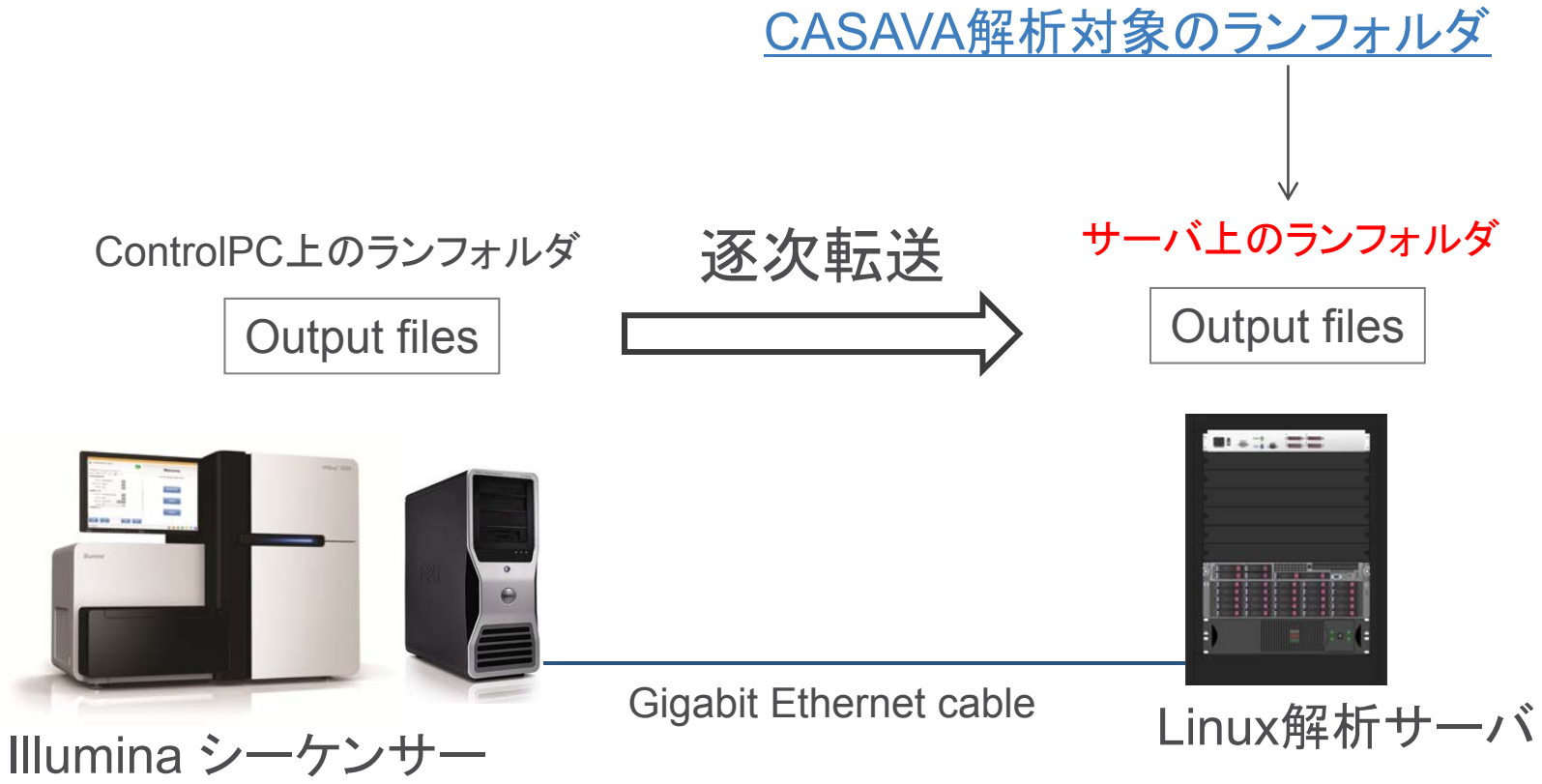
illumina®

illumina®

シーケンサーから解析場所へのデータ転送

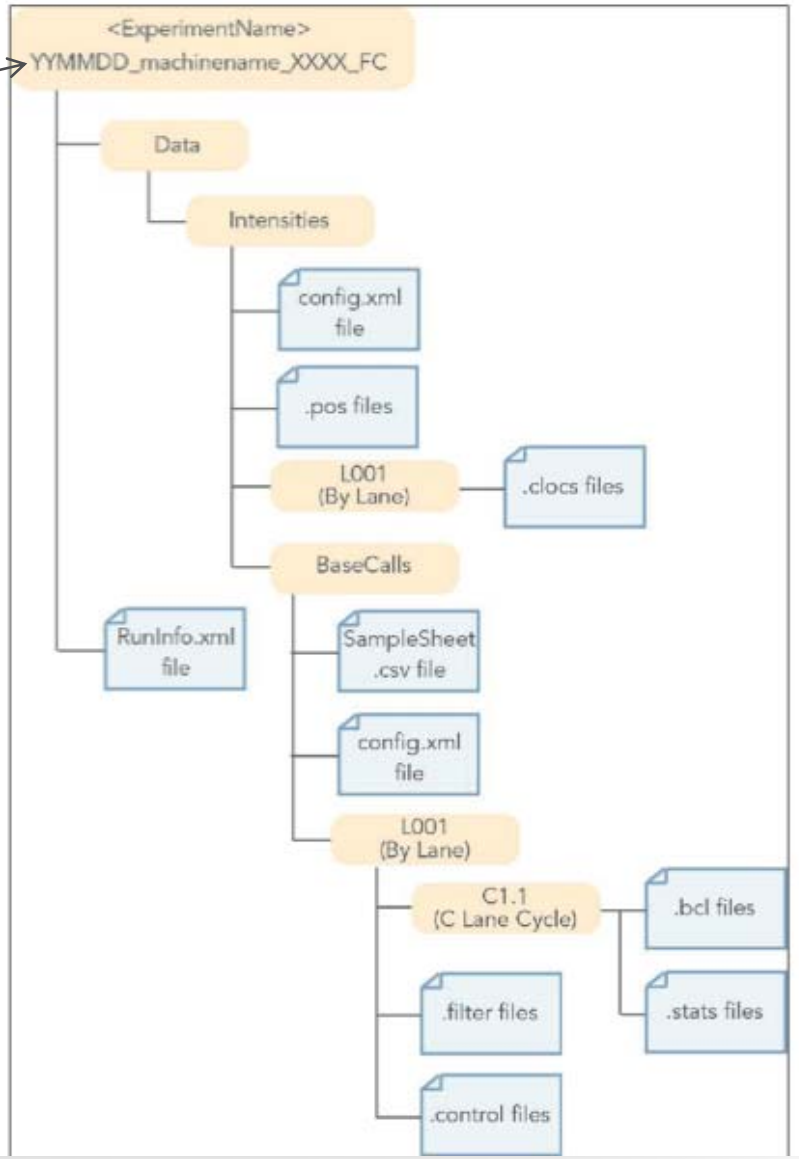
HiSeq/GA では、シーケンス時にデータをサーバに逐次転送する。

(* MiSeqはデフォルトでは内蔵ディスクに蓄積)



ランフォルダ構造

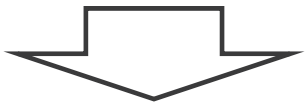
ランフォルダ



illumina®

CASAVAでのFASTQ生成に必要な主要ファイル

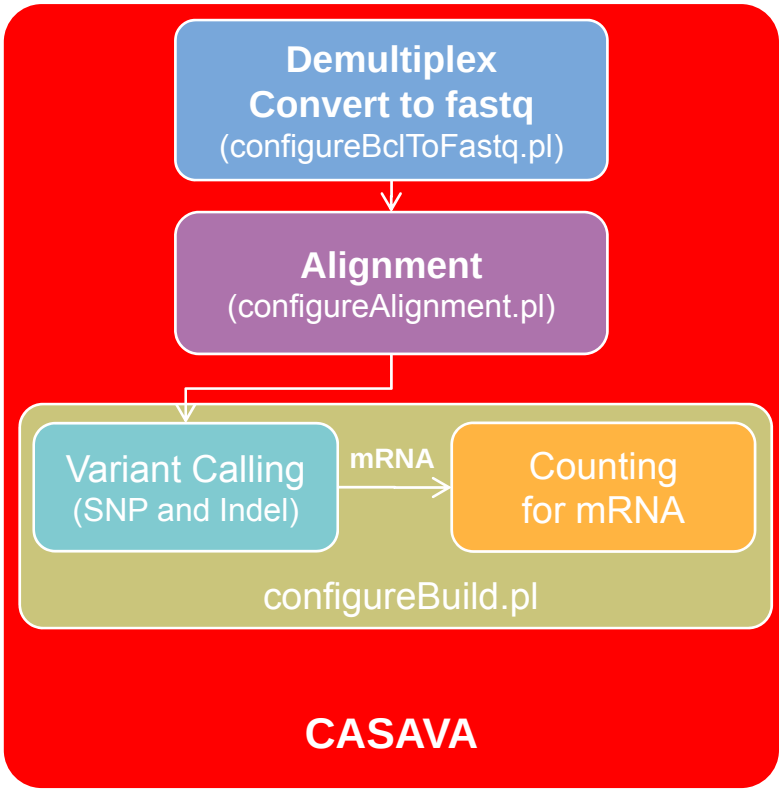
- .clocs ファイル、.locsファイル、posファイル：クラスタの位置情報
- .bcl ファイル:配列の情報が入ったデータ
- .filter ファイル:Pass Filter が通ったかどうかの情報
- .stats ファイル:平均蛍光強度などの統計情報
- .control ファイル:cross-talk matrix の作成レーンか、PhiXにアライメントできたか、In-line controlか、などのコントロールに関わる情報
- RunInfo.xml
- config.xml



これらを含み、かつディレクトリの階層構造を保ったままのランフォルダがCASAVAでは必要となる

CASAVA v1.8.2 User Guide p.27

CASAVAの処理体系



CASAVA v1.8 の主要3コマンド

本日お話しする範囲

① configureBclToFastq.pl ・・・FASTQ生成 結果：Unaligned/ 配下, fastq.gz

cd Unaligned
make

② configureAlignment.pl ・・・アライメントを実行 結果：Aligned/ 配下, .bam

cd Aligned
make

③ configureBuild.pl ・・・変異コールを実行 結果：Build/ 配下, .vcf

cd Build
make

- CASAVAで行いたい解析まで順番に実行する。
- どのコマンドも、ランフォルダ全体が必要。
- HiSeq/GAは少なくとも configureBclToFastq.pl まではCASAVAが必要となる。

(*クラウド環境 BaseSpace ご利用の場合を除く)

configureBclToFastq.pl の実行

configureBclToFastq.pl 概要

- CASAVA1.8 の configureBclToFastq.pl コマンド

  FASTQへの変換とサンプル毎への配列振分けの両方を一度に処理

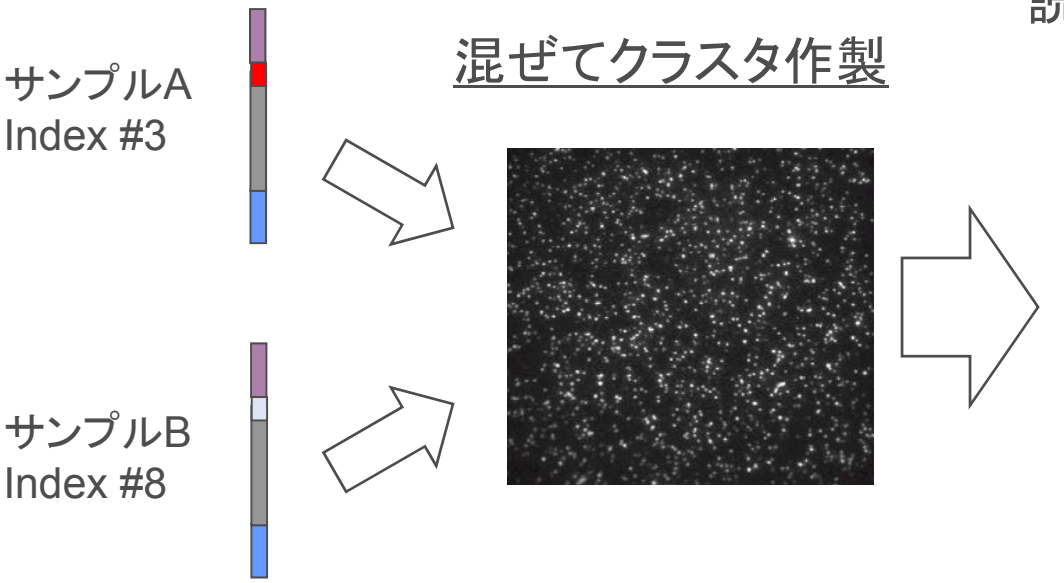
  – 装置の出力結果を、業界標準のFASTQフォーマットに変換
  – 複数混ぜたサンプル(Multiplex)を、指定されたサンプル毎にまとめる
    ->デマルチプレクシングと呼称

- Multiplexした場合は、SampleSheet.csv という指示書のような

  カンマ区切りファイルの作成が必要

Multiplexing のラン、インデクス配列

- 1レーンに複数サンプルを流したラン
- サンプル毎に予め異なるindex配列（バーコード）をもつアダプタをサンプルDNAにつけることにより、複数サンプルを混ぜてランを行っても、後の工程でindex配列を文字列としてソフトウェア的に認識することで、どのサンプルのリードであったか判別できる
- この配列の振分けのことを、Demultiplex（デマルチプレックス）という



読まれた配列の例

リード配列：ACCTGATCTGTATG…
Index配列：**TTAGGC**
→ **index #3**が見つかったので
サンプルAだとわかる

リード配列：ACCTGATCTGTATG…
Index配列：**ACTTGA**
→ **index #8**が見つかったので
サンプルBだとわかる

illumina

configureBclToFastq.pl の使い方 (Linux コマンド)

1. cd /path/to/ランフォルダ名

解析したいランフォルダの場所に移動

2. /illumina/software/CASAVA-1.8.x/bin/configureBclToFastq.pl --input-dir /path/to/run/data/Intensities/BaseCalls --sample-sheet /path/to/SampleSheet.csv

必要ファイルのチェックと
実行の場となる Unalignedディレクトリの作成

3. cd /path/to/run/Unaligned

Unalignedディレクトリへ移動

4. nohup make -j <n>

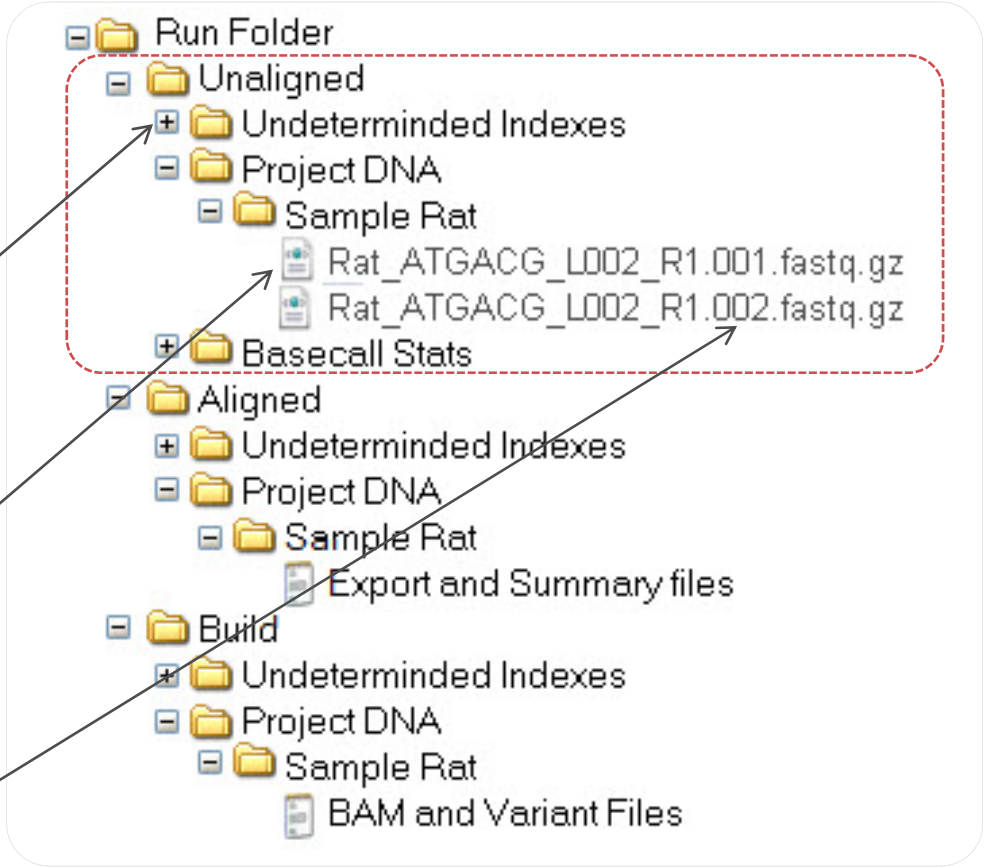
実行

- コマンドは1行として打ち込む(改行は含まれません)
- ランのディレクトリは上記の例で /path/to/run/ になっているので、実際のディレクトリに置き換える
- ランフォルダ名は解析したい英数字のランフォルダ名に置き換える
- 上記の例では /illumina/software/CASAVA-1.8.x/ にCASAVAがインストールされた設定になっているので実際は使用環境のインストール先を指定する
- Linuxではフォルダのことをディレクトリと呼ぶ
- <n> は使うCPU 数なので任意の数値に置き換える('<'と'>'の記号も取り除く)

illumina®

BCL からFASTQへの変換と Demultiplexing 出力ファイル

- SampleSheet.csv に基づきサンプルごとにレーンごと、インデックスごとに記載される。
- Basecall Stats フォルダはランの結果のsummaryを含む
- SampleSheet で指定したどのindex配列にも該当しなかった配列は、Undetermined_Indices/ ディレクトリ配下に,レーン毎のまとまりでFASTQとして生成される
- FASTQファイルは圧縮されている
- デフォルトでは400万リード毎に別のインクリメントされた名前のファイルが作成される



configureBclToFastq.pl の主要なオプション

オプション	内容
--input-dir /pathto/Data/Intensities/BaseCalls	BaseCalls のディレクトリの指定
--output-dir /path/to/output	出力ディレクトリの指定
--sample-sheet /path/to/Sample.csv	SampleSheet.csvの指定
--fastq-cluster-count N	fastqファイルに含まれるクラスタの数を入れる デフォルトは4,000,000。 N=0とすると1つのfastqとしてまとまる （ただし、サンプル毎のリード毎）。
--use-bases-mask Y35n,I6n,Y35n	どの塩基を使うかを指定する
--mismatches N	許可する index のミスマッチの数デフォルトは0 (0,1,2を指定可能)
--with-failed-reads	Pass Filter に通らなかったリード配列も fastq に出力する

※ N は指定したい数字で置換

--use-bases-mask オプション　:使用塩基のマスク

- 記号の意味について
  - Y　通常のリードとして使用する塩基
  - n　使用しない塩基
  - I (あい)　インデックス塩基
  - 数字　指定した塩基数（サイクル数）分、一つ前の記号として続ける
- 例
  - Y36　→　36塩基を使う
  - Y35n　→　最初の35塩基を使い、最後の塩基を使わないようマスクする
  - Y35n,Y35n　→　Read1と2の最初の35塩基を使い最後の塩基をマスクする
  - Y35n,I6n,Y35n　→　Read1と2の最初の35塩基を使い最後の塩基をマスクして、インデックス配列を6塩基使う（ここでは、インデックス配列は7塩基読んでいるものの最後の塩基は使わないと指定している）
  - Y*,I*,I*,Y* →　Read1と2を全て使い、dual indexを使用したがこれも全て使う

* 全体として(Y,n,Iを合わせて)、シーケンスしたサイクル数（塩基長さ）と整合させる必要がある

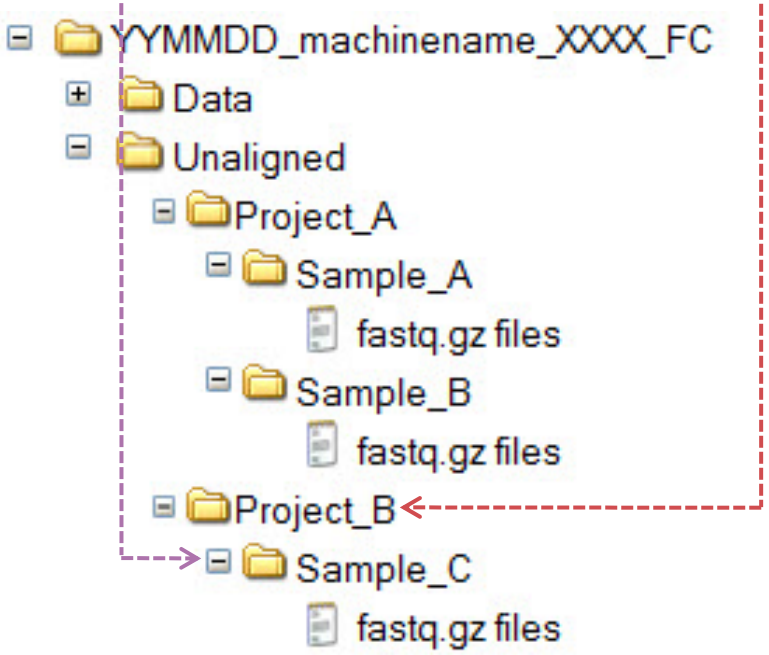
コマンド操作 例:レーン毎のFASTQを得る
PE, インデクス無し

SampleSheet.csv

SampleSheet.csv とフォルダ構造

	A	B	C	D	E	F	G	H	I	J
1	FCID	Lane	SampleID	SampleRef	Index	Description	Control	Recipe	Operator	SampleProject
2	FC200DMAB	2	A	hg18	ATCACG	Example	N	PE_Indexing	FZ	A
3	FC200DMAB	2	B	hg18	CGATGT	Example	N	PE_Indexing	FZ	A
4	FC200DMAB	3	C	hg18	ATCACG	Example	N	PE_Indexing	FZ	B

Sheet1 Sheet2 Sheet3



*FASTQが指定の単位にまとまる

*後続のAlignmentやVariant Call解析をCASAVAで続ける場合も、FASTQ生成以降は基本的にこのProject xSampleの単位で実施される

illumina®

illumina®

SampleSheet.csv の各列について

- FCID　フローセルID（省略可）
- **Lane**　レーン
- **SampleID**　サンプルID
- SampleRef　サンプルのリファレンス（省略可）
- Index　インデックス配列（インデクスを使用していない場合省略可）
- Description　コメント（省略可）
- Control　コントロールレーンかどうか（省略可）
- Recipe　Sequencing に使用したレシピ（省略可）
- Operator　Sequencing を実施した人（省略可）
- **SampleProject**　プロジェクト名（省略可）

• **太字の項目は省略しないことを推奨**

illumina®

illumina®

SampleSheetの作成例方法

- SampleSheet は、**カンマ区切りテキストファイル**
  - それぞれの行ごとに、同数のカンマで区切られた項目がある
- Excel などでテーブルを作製し、.csvファイル形式として保存したものを SampleSheet として使うことができる
- Notepadなどテキストエディタで開いて確認すると以下例のように見える
- 作成後は、ランフォルダの直下に配置することをお勧めしている

例；

```
FCID,Lane,SampleID,SampleRef,Index,Description,Control,Recipe,Operator,Project
FC626BWAAXX,4,SampleA,Mouse,ATCACG,'DefaultSample',N,,,ProjectX
FC626BWAAXX,4,SampleB,Mouse,CGATGT,'DefaultSample',N,,,ProjectX
…
```

- CASAVAのサンプルシートは、MSRのものと書式がやや異なる。
- MiSeqのランフォルダを処理する場合、CASAVA様式のサンプルシートをまず用意する必要がある。

Sample Sheetをエクセル等で表示した例

Sample Name **Barcode/Index** **Project Name**

FCID	Lane	**Sample**	SampleRef	**Index**	Description	Control	Recipe	Operator	**Project**
FC62DBU	1	NA12156_Index_1	Human	ATCACG	1:250method1	N	test.xml	AT	testProject1
FC62DBU	1	NA11992_Index_2	Human	CGATGT	1:250method1	N	test.xml	AT	testProject1
FC62DBU	1	NA11882_Index_4	Human	TGACCA	1:250method1	N	test.xml	AT	testProject1
FC62DBU	1	NA11881_Index_3	Human	TTAGGC	1:250method1	N	test.xml	AT	testProject1
FC62DBU	1	lane1	unknown	Undetermined	unknown barcode	N	test.xml	AT	Undetermined_indices
FC62DBU	2	NA12156_Index_1	Human	ATCACG	1:500method1	N	test.xml	AT	testProject1
FC62DBU	2	NA11992_Index_2	Human	CGATGT	1:500method1	N	test.xml	AT	testProject1
FC62DBU	2	NA11882_Index_4	Human	TGACCA	1:500method1	N	test.xml	AT	testProject1
FC62DBU	2	NA11881_Index_3	Human	TTAGGC	1:500method1	N	test.xml	AT	testProject1
FC62DBU	2	lane2	unknown	Undetermined	unknown barcode	N	test.xml	AT	Undetermined_indices
FC62DBU	3	NA12156_Index_1	Human	ATCACG	equal Volume	N	test.xml	AT	testProject1
FC62DBU	3	NA11992_Index_2	Human	CGATGT	equal Volume	N	test.xml	AT	testProject1
FC62DBU	3	NA11882_Index_4	Human	TGACCA	equal Volume	N	test.xml	AT	testProject1
FC62DBU	3	NA11881_Index_3	Human	TTAGGC	equal Volume	N	test.xml	AT	testProject1
FC62DBU	3	lane3	unknown	Undetermined	unknown barcode	N	test.xml	AT	Undetermined_indices

- CASAVAのサンプルシートは、MSRのものと書式が異なる
- Notepadで開くと前頁のように同数のカンマで区切られているもの

禁忌文字

Illegal Characters

? ( ) [ ] / \ = + < > : ; “ ‘ , * ^ | & .　とスペース、全角文字

* Alignment以降も実行の場合、カスタムFASTAのコメント部も上記が同様に禁忌となります。

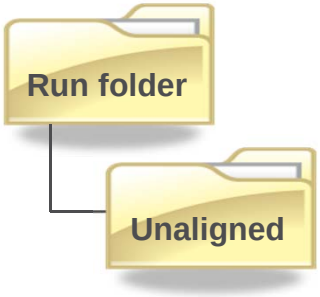
*CASAVA v1.8.2 User Guide p.31

コマンド操作デモ：　複数サンプルをデマルチプレックスし、
サンプル毎にFASTQを得る

MiSeqランフォルダ,SE, single index の例

‘Project-based’ Run Folder

The Sample Sheet provides names of Project and Samples



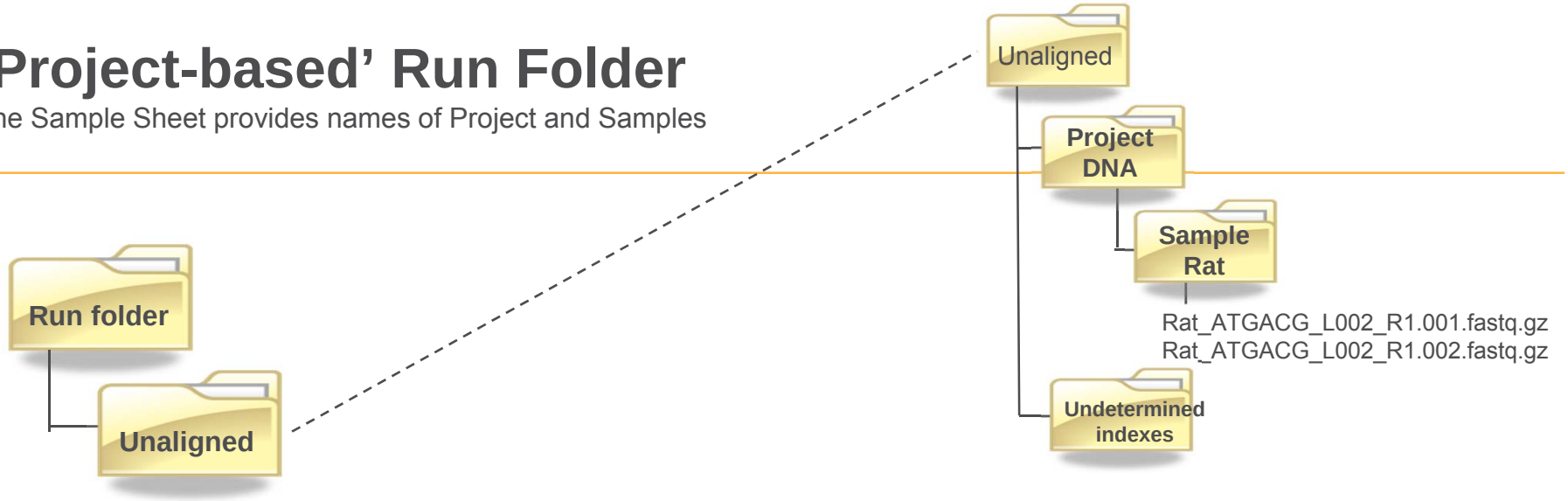
Bcl conversion and
De-multiplexing

illumina®

illumina®

‘Project-based’ Run Folder

The Sample Sheet provides names of Project and Samples

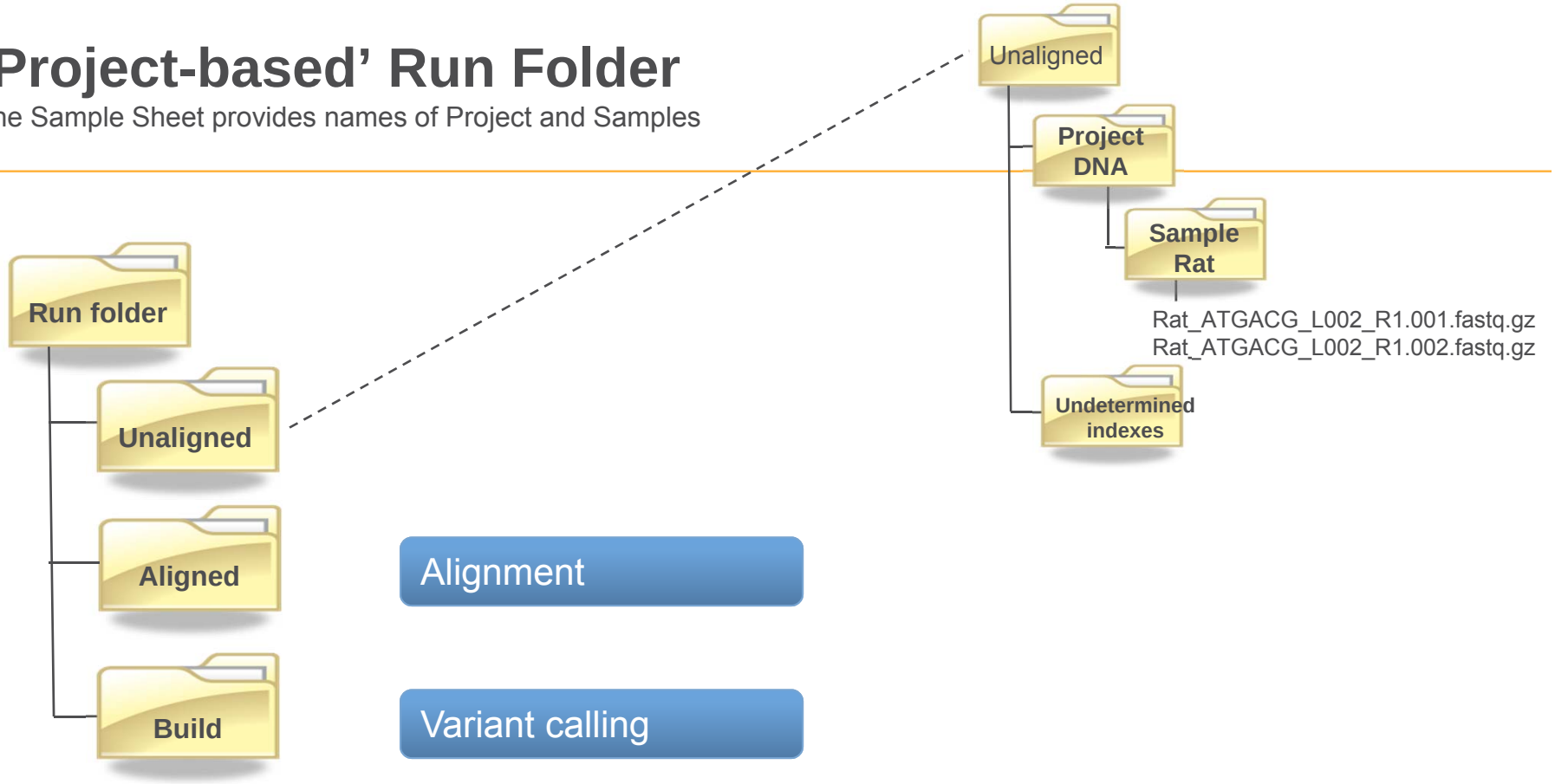


illumina®

illumina®

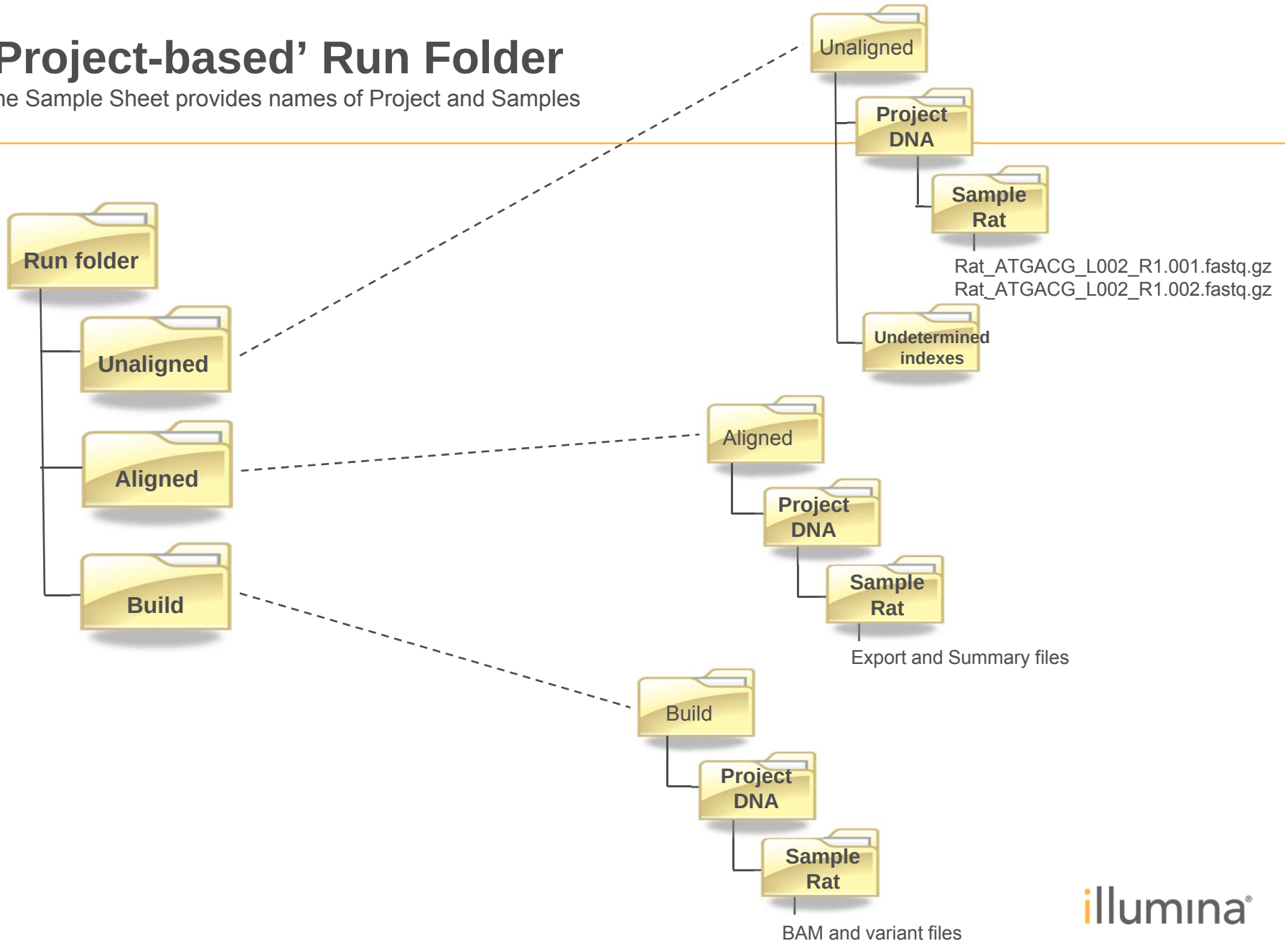
‘Project-based’ Run Folder

The Sample Sheet provides names of Project and Samples



‘Project-based’ Run Folder

The Sample Sheet provides names of Project and Samples



デマルチプレックス時のindexの扱いにおける MSRとCASAVAの主な違い

CASAVA	MSR
デフォルトmismatches 許容は0	デフォルトmismatches許容は1
mismatchは0|1|2を指定できる	mismatchは固定で変更できない
--sample-sheetで指定するため サンプルシート名は可変	サンプルシート名はSampleSheet.csvで固定。 ファイル配置位置とファイル名で認識される
--use-bases-mask の指定で詳細に使用塩基のマスクが可能	基本的には使用塩基を選べない
デフォルトではI*nとして最後の塩基のマスクを想定する。即ち、サンプルシートに記入のインデクス塩基数は、実際読んだ塩基数からマイナス１した数が記入されていることを前提としている 全ての塩基を使用する場合は明示する必要がある dual index利用の場合も明示する必要がある	デフォルトではすべての塩基を使用 実際にシーケンスしたインデクスサイクル数分の塩基をサンプルシートに記入する

※ N は指定したい数字に置き換える

コマンド操作例：　複数サンプルをデマルチプレックスして
サンプル毎にFASTQを得る

dual index の例

CASAVAでdual indexを処理する方法

- --use-bases-maskで明示的に指示する必要がある
- サンプルシートのインデクス配列の列には、ハイフンでつなげて記入する

Sample Sheets

Index Type	Library Type	Index Length
Single index	TruSeq LT/v2 TruSeq Small RNA	Seven-base Index Read; only six bases are used for demultiplexing Example: CGATGT
Dual index	TruSeq HT Nextera	Eight-base Index Read Example: TAGATCGC-CGTACTAG

Demultiplexing

Run CASAVA 1.8.2 separately for each index type and sample sheet using the -- use-bases-mask command for dual-indexed runs to define how each sequenced read should be used. This command is required to demultiplex for either a six-base or eight-base index.

The following --use-bases-mask examples represent paired-end runs. For more information, see the *CASAVA v1.8.2 User Guide*, Part # 15011196.

Index Type	--use-bases-mask	Description
Single index	Y*, I6n*, n*, Y*	Mask for Index 1 and only six bases
Dual index	Y*, I8, I8, Y*	Mask for dual indices and all eight bases

*「Sequencing Mixed Libraries on a HiSeq or GA Flow Cell」

サンプルシートを用いている場合は、記入した使用インデクス配列数と、
--use-bases-maskで指示した塩基が整合するよう注意が必用である。

コマンド操作例： インデクス配列をFASTQとして出力する

インデクス配列をFASTQとして生成する

トラブルシュート等で必要となった場合の処置として、
--use-bases-maskで明示的にＩ(インデクス塩基指定)をY(通常リード塩基)として
指示することでインデクスFASTQを生成することができる

例；

--use-bases-mask=Y*,I*,Y*	->	--use-bases-mask=Y*,Y*,Y*
--use-bases-mask=Y*,I*,I*,Y*	->	--use-bases-mask=Y*,Y*,Y*,Y*
--use-bases-mask=Y*,I*	->	--use-bases-mask=Y*,Y*,Y*
--use-bases-mask=Y*,I*,I*	->	--use-bases-mask=Y*,Y*,Y*

illumina®

illumina®

bcl2fastq v1.8.4　ソフトウェア

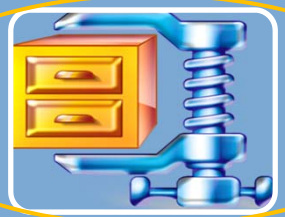
*HiSeq 1500/2500でデータ圧縮機能利用時のみ関係します

HiSeq Data Storage Options

BaseSpace へのランデータのアップロード

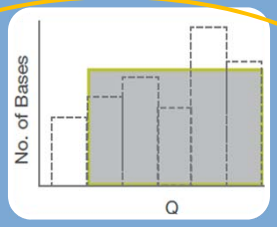
- 保存のためにリアルタイムで BaseSpace へデータを転送

BCL files* を圧縮



- ランに必要なディスクスペースを減少

Qscores* のバイナリ化



- ランに必要なディスクスペースを減少
- 一定の範囲の Q scores をグループ化

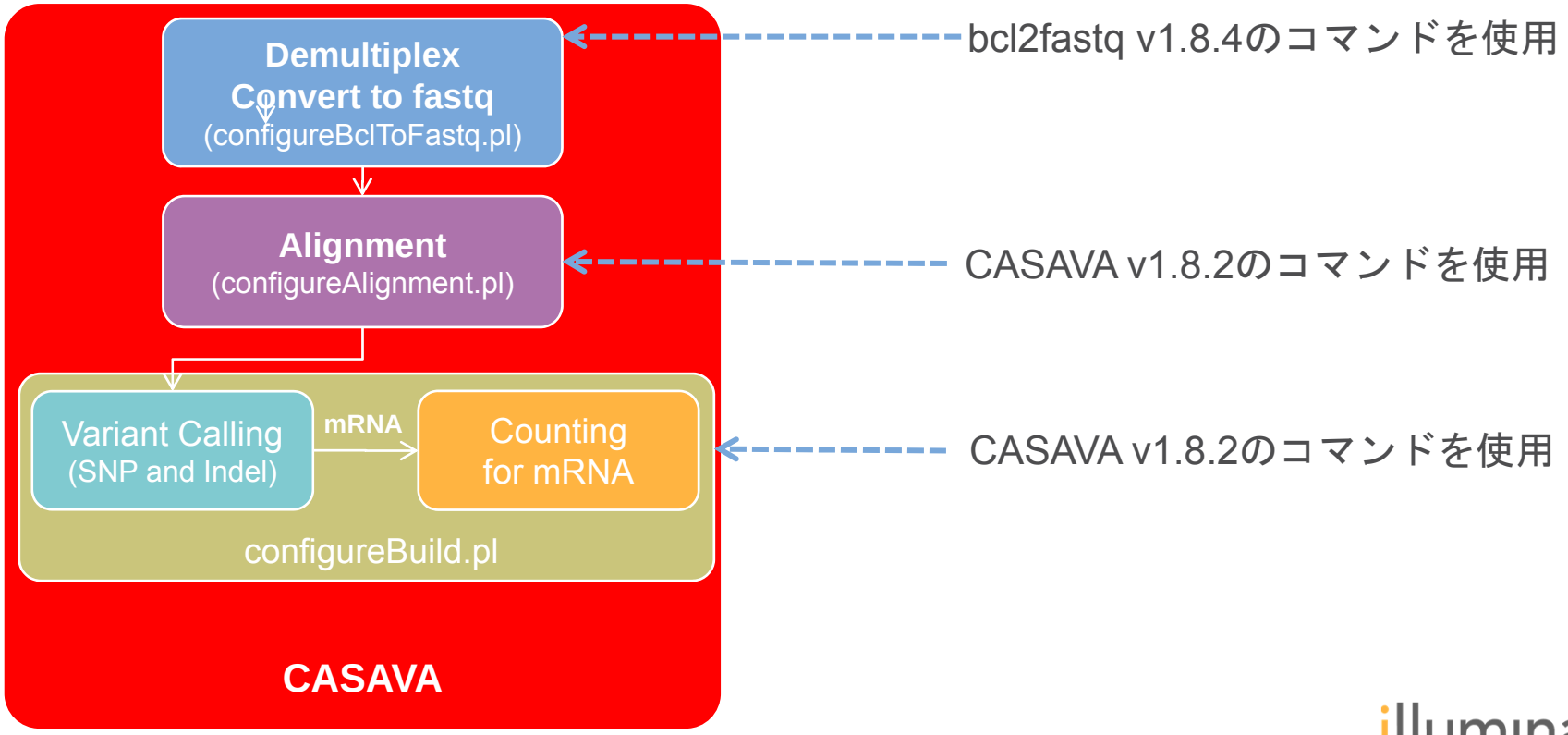
*詳細情報は、ホワイトペーパー「*Reducing Whole-Genome Data Storage Footprint*」を参照

を選択した場合 configureBclToFastq.plのv1.8.4が必用です

bcl2fastq v1.8.4 が提供するconfigureBclToFastq.pl (v1.8.4)

*HiSeq 1500/2500でデータ圧縮機能利用時

- HiSeqでシーケンスを開始時にbcl zipping やQscore binningの使用を指定された場合は、v1.8.4のconfigureBclToFastq.pl を使用する必要がある。パッケージはCASAVAとしてではなく、bcl2fastq ソフトウェアとしてconfigureBclToFastq.pl部分のみを配布している。



トラブルシュート時にまずお送り頂きたいファイル

- ランフォルダ/SampleSheet.csv　(お使いのsamplesheet.csv)
- ランフォルダ/Unaligned/support.txt
- ランフォルダ/Unaligned/nohup.out
- ランフォルダ/RunParameters.xml

ご参考資料

- CASAVA v1.8.2 User Guide (英語)
  - インストール方法を含みます

  http://support.illumina.com/sequencing/sequencing_software/casava.ilmn
  から、左ペインの Documentation & Literature へ

- CASAVA v1.8.2 PhiXを用いたハンズオン (日本語, 非公式補足資料)
  - テクニカルサポートまでお問合せください。

- オンラインコース(英語)

http://support.illumina.com/sequencing/sequencing_software/casava/training.ilmn

- Q&A(英語)

http://support.illumina.com/sequencing/sequencing_software/casava/questions.ilmn

illumina®

illumina®