2014年11月14日
サポートウェビナー

NextSeq 500から得られるデータのFASTQ変換

- bcl2fastq バージョン2 ほか

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)

illumina®

本日の内容

- NextSeq 出力データを扱うための３選択
  - 選択１：BaseSpace クラウドを使用する
  - 選択２：BaseSpace オンサイトを使用する
  - 選択３：ご自分の計算機環境を使用する

- bcl2fastq2の使い方 (NextSeq 500, HiSeq X Ten 共通)
  - bcl2fastq2 とは
  - サンプルシートの準備
  - 実行前後のファイル構造とNextSeqデータ圧縮
  - Bcl2fastq2の実行
  - 実行結果レポート

サンプルから答えまでのワークフロー

MCS/HCS/NCS/RTA
ソフトウェア
on ControlPC

画像取得 /シグナル抽出

↓

ベースコール (*.bcl)

シーケンシング中

↓

*.bcl ファイルから **FASTQに変換**

bcl2fastq (Linux)
bcl2fastq2 (Linux)
MiSeq Reporter (Win)

↓

再解析、その他下流解析、可視化、
アノテーション、フィルタリング
レポート生成など

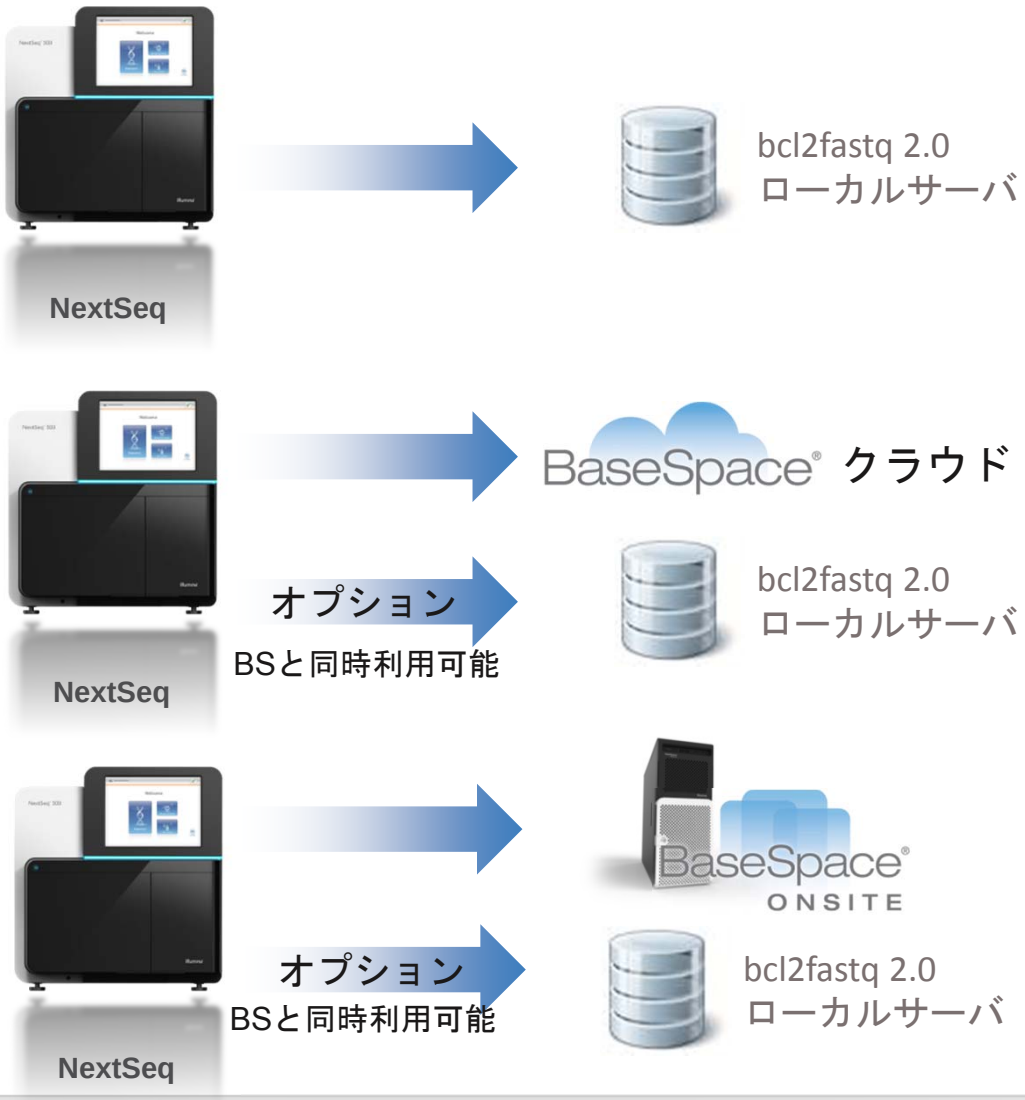
シーケンシング後

二次解析以降

ワークフロー化

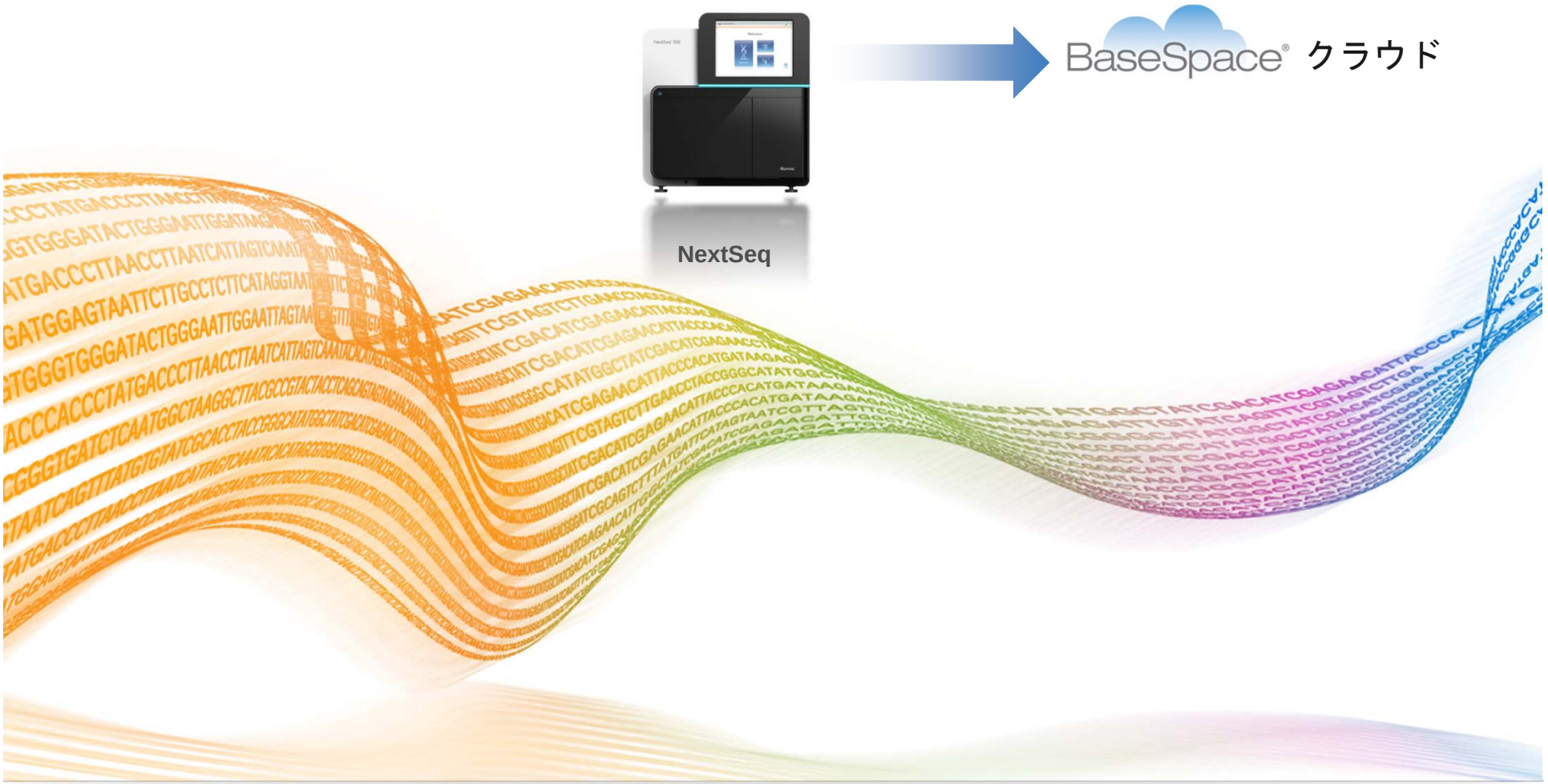
illumina

NextSeq データを扱うための３選択 サマリ



illumina

選択１：
BaseSpace クラウドを利用する



illumina®

BaseSpaceはどなたでも、すぐ使い始められます。

- メールアドレスを持つ人ならどなたでも、ほぼマウスクリックの操作のみで解析や情報管理・共有を行うことができる環境です。
- クラウド利用タイプと、ローカルサーバタイプでご提供いたします。
- イルミナ製品をお持ちの方も、お持ちでない方も、メールアドレスでアカウント登録頂ければ使えます。

basespace.com からログイン頂ければ公開デモデータを使ってすぐに解析をお試し頂けます。

cf. http://aws.amazon.com/security

- ”Amazon Web Services” を土台のクラウド環境として使っています。
- **ディスク1Tバイトの容量までフリーの利用（1アカウントあたり）**
  イルミナ コア アプリもフリーの利用

(RNA Seq，エクソーム、全ゲノム、腫瘍/正常解析(全ゲノム)、16S メタゲノム解析、VariantStudio アノテーション・変異解析ツールなど)

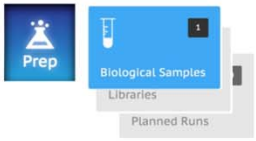
BaseSpace クラウド：サンプルから答えまで

インターネットがあれば初期費用ゼロでスタートできます

NeoPrep,
サンプル管理

BaseSpace® BaseSpace® BaseSpace®

Library Prep | Sequence | Analyze Data | View Results | Collaborate



DNA, RNA
& ターゲティッド
ソリューション

リアルタイムモニタリング
過去のクオリティ値等の比較管理

データQC、
イルミナ無償アプリと
多彩な他社製アプリによる解析

多彩なアノテーションや視覚化

解析管理、
共有、データ保全計画

サンプル情報管理から解析の繰返しや比較、レポーティングまでを簡単に一元管理

illumina®

クラウドご利用イメージ例；

10Mbps

BaseSpace™

クラウド

Gbイーサ

ローカルサーバ

ラン終了４分程度後には、以下<u>２セット</u>のデータがローカルとクラウド上に用意されている

- クラウド上のランフォルダ
  転送後自動でdemultiplexingが実施される
- 転送先ローカルサーバ上 のランフォルダ
  bcl2fastq2を実施頂きFASTQ生成を実施可能

2セットのデータを
物理的に別々の場所に保有

- **Data back up**
- **Robust system**
- **Disaster planning**

illumina®

BaseSpace クラウドへのシーケンサー 接続要件

インターネットがNextSeqから利用できる環境であること
-> NextSeqのWebブラウザから、Googleなどが見られる状況で
したらOK

＜所内ファイヤーウォールの設定等ある方向け情報＞

BaseSpace利用時の使用ポート；
ポート80, 443 (HTTPS/SSL)
-> 一般的なwebブラウザで使用されているポートです

また以下の名前でのインターネットへのアクセス；
api.basespace.illumina.com
basespace.illumina.com
*.amazonaws.com

NextSeq BaseSpaceクラウド ご参考

クラウドご利用にご興味いただけた方は、弊社HPに加え、
イルミナサポートウェビナー もぜひご参考下さい。
http://www.illuminakk.co.jp/events/webinar_japan.ilmn?ws=ss

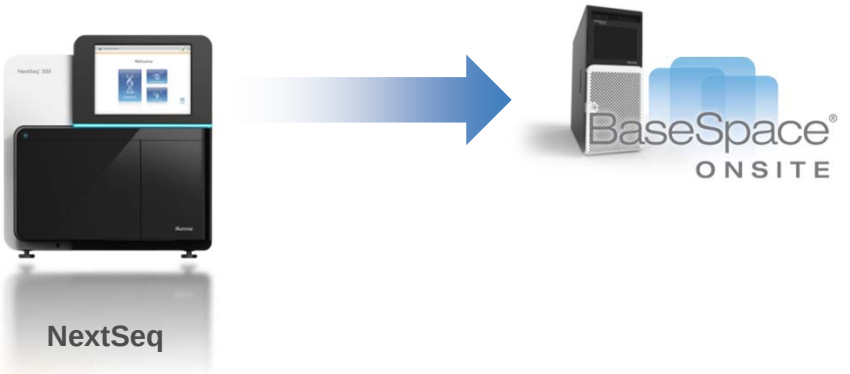
BaseSpace 環境自体の使い方に関するウェビナー

2014/09/12 **「次世代シーケンサー（NGS）の新たなデータ解析アプローチ：BaseSpace 」**

2013/09/06 サポートウェビナーシリーズ 2013
「BaseSpace リリース版（MiSeq/HiSeq）」

2012/11/22 サポートウェビナーシリーズ 2012
「BaseSpace - genomics cloud computing -- ベーススペースの使いかた」

選択２：
BaseSpace Onsiteを利用する



illumina

BaseSpace Onsite (ベーススペースオンサイト)

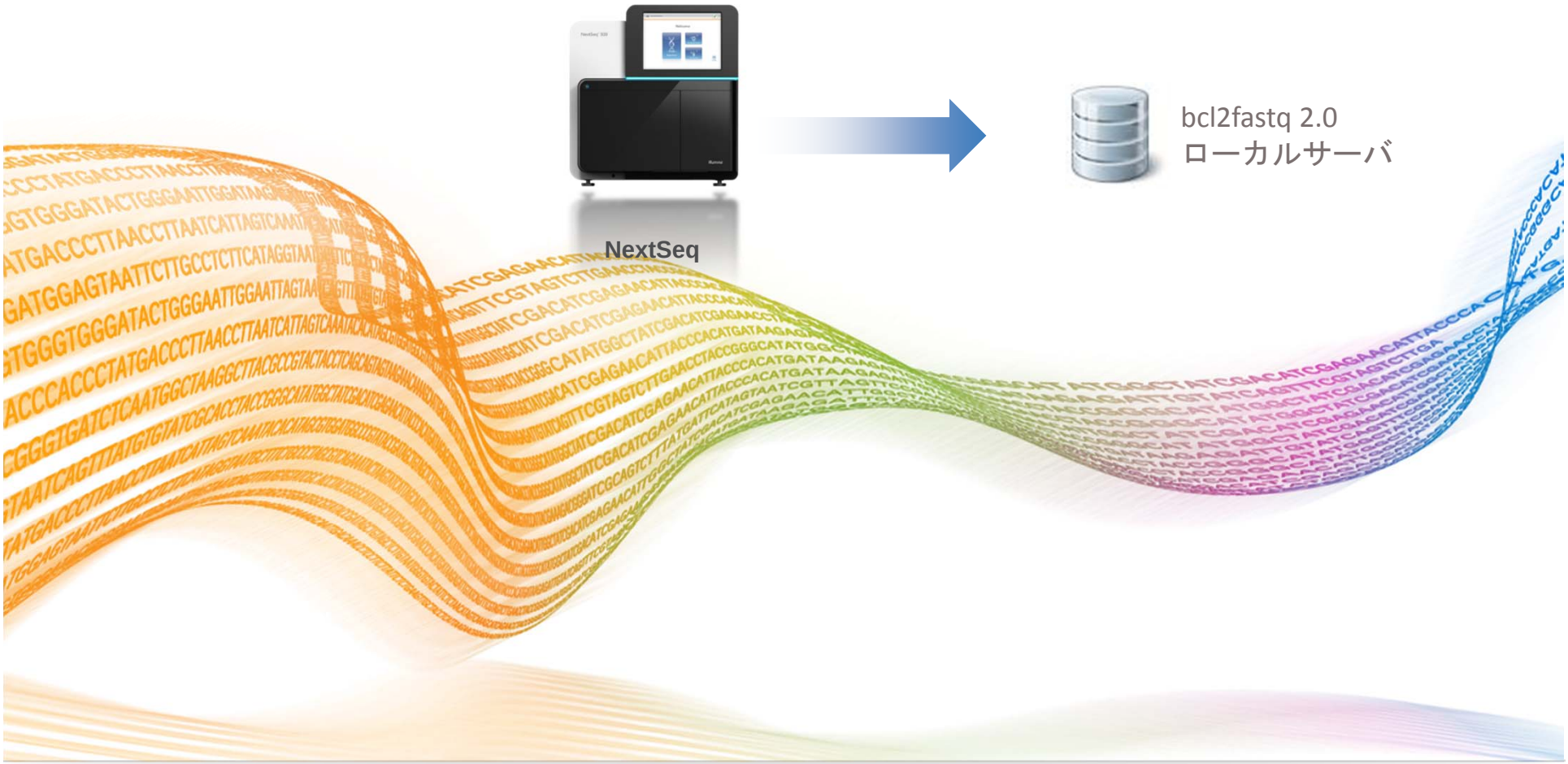
外部インターネットに接続しないローカルサーバと解析環境

- データはオンサイト内で保管、インターネットの接続を必要としません
- 所内のイントラネットに接続して利用
- 現在NextSeqに対応（HiSeq/MiSeqにも2015 対応予定）
- イルミナによるトータルサポート
- サンプル、ラン、解析データを一元管理； ITやバイオインフォの負荷を軽減
- BaseSpaceクラウドと同様の使い勝手
- 他社製アプリは現在未搭載
- イルミナコアアプリのみを搭載
- アップデートによりアプリ搭載数も増加予定
- 最大６ノード構成

BaseSpace Onsite 価格

カタログ番号	製品	内容	価格
SE-403-1001	BaseSpace Onsite システム	• BaseSpace Onsite 4U サーバ • BaseSpace Onsite システムソフトウェア • **1 年間のサービスとSWライセンス込み**	900万円
SW-430-1001	年間サービス +SW ライセンス	• 機能やユーザーインターフェースの改善 • ハードウェアサービス	225万円
SE-403-1004	ブリッジ	• 2台以上のBaseSpace Onsiteシステムへのコネクション変更	0 円

選択３：
ご自分の計算機環境を使用する



illumina®

シーケンス中のデータ転送先サーバのご準備

- NextSeqは塩基をコールする度に逐次データをサーバに転送する
- このため**転送先サーバの準備が必用**
- NextSeq 内蔵のWindowsPC からサーバに書き込める状態を構築することが必用（Samba などCIFSプロトコルによる共有環境の構築）

NextSeq 内で
ベースコール

シーケンス中は断続的に逐次自動転送がつづく

LANケーブル (Gbイーサのみサポート)

サーバ上に出力フォルダができ、中にデータが蓄積されていく

ご自分のサーバ
でFASTQに手動変換

illumina

NextSeq データ量 (スループット)

	デスクトップ型		大型	
	フォーカス	柔軟	生産性	集団規模
				
	MiSeq	NextSeq	HiSeq 2500	HiSeq X Ten
ランあたりの最大データ量	15 Gb	120 Gb	1000 Gb	1800 Gb
最大リード長	300x2	150x2	250x2	150x2
				(全ゲノムのみ)

※ランのリード長（サイクル長）、Mode、PE/SR、クラスタ密度等により多少変わって参ります。

シーケンス中のデータ転送の速度要件

高出力フローセル (HO) の場合で**約120GByte 程度を29時間** でシーケンス

↓

このため10Mbps 以上を推奨

実際はカタログ値の120Gbases よりも多くデータが出てしまう事も多く、
またピーク速度も考慮して、この数倍の転送速度が確保できていれば
より安心

NextSeq 500 サイトプレップ（事前準備）ガイドにも記載がございます。
公式な要件ドキュメントが必要の場合はサイトプレップガイドをご参照ください。

Linux コマンドラインツール bcl2fastq2

要件を満たす計算機環境をお持ちであれば、
弊社より無償で提供のbcl2fastq2をインストールし、利用できる。

Bcl2fastq2はFASTQ生成ツールであり、解析機能はない。

ご自分のサーバ環境構築・運用・トラブルシュート、
ソフトウェアインストール・アップグレード等は
お客様でのご準備・ご実施となる。

弊社製品に関してはテクニカルサポートからのアドバイスは随時得られる。

NextSeq の場合の３選択　まとめ

弊社独自フォーマットのbcl形式を業界標準FASTQ形式に変換するため、
どれか１つは必ずご用意頂く必要がございます（併用運用可能）。

BCL ファイル → FASTQ 作成 → 解析（アライメント など） → レポート 作成

お持ちのLinux サーバ利用	Linuxソフト bclfastq2	オープンソース & 有償ツール	• コマンド打ち込み必要 • バイオインフォマティクス、IT管理必要
BaseSpace Cloud （インターネット経由）	ウェブブラウザーで操作		• コマンド不要 • インターネット必要 • 初期費用不要
BaseSpace Onsite (ローカルサーバ)	ウェブブラウザーで操作		• コマンド不要 • インターネット不要 • イントラネット必用 • 購入費用必要

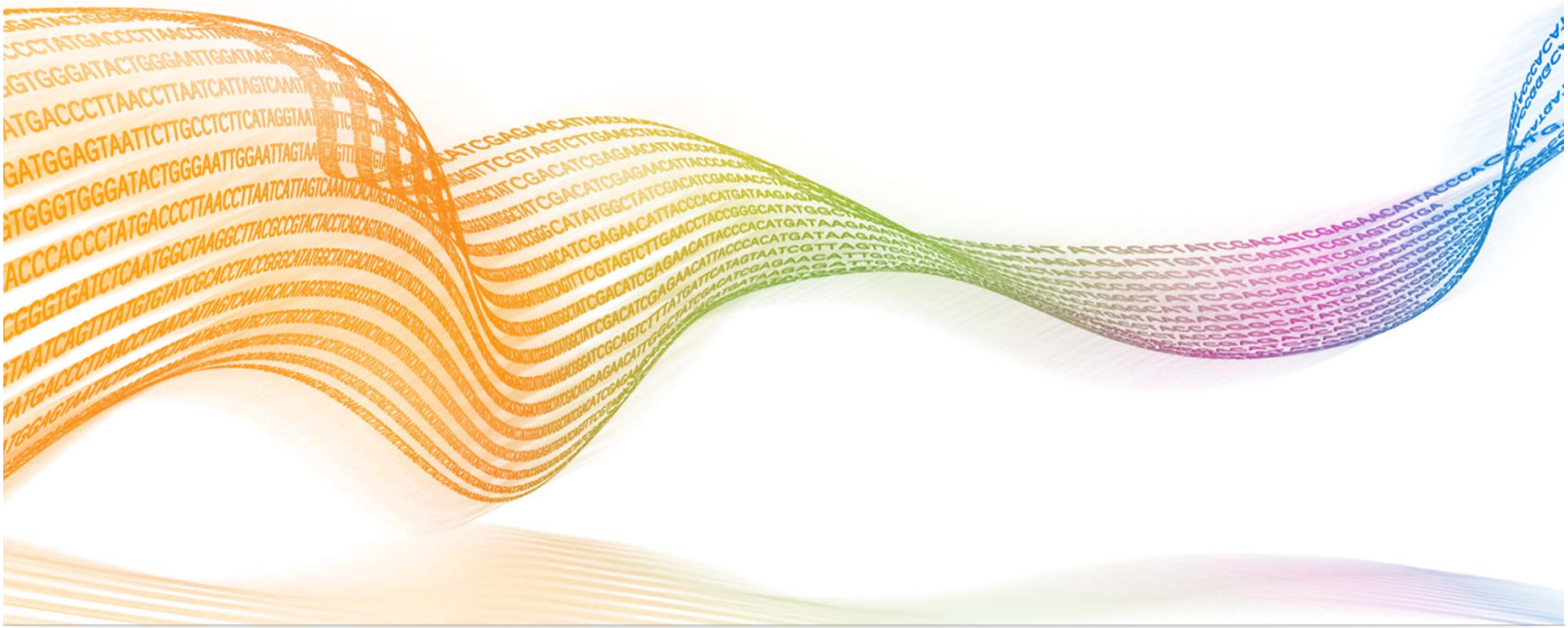
illumina

本日の内容

- NextSeq 出力データを扱うための３選択
  - 選択１：BaseSpace クラウドを使用する
  - 選択２：BaseSpace オンサイトを使用する
  - 選択３：ご自分の計算機環境を使用する

- bcl2fastq2の使い方 (NextSeq 500, HiSeq X Ten 共通)
  - bcl2fastq2 とは
  - サンプルシートの準備
  - 実行前後のファイル構造とNextSeqデータ圧縮
  - Bcl2fastq2の実行
  - 実行結果レポート

bcl2fastq2とは



illumina®

シーケンス終了後のFASTQ変換用 Linuxコマンドラインツール：bcl2fastq2

- ベースコールファイル(*.bcl類)を FASTQ に変換する
- Linux にインストールしてコマンド1行程度を打込み実行
- ソフトウェアは無償
- ソフトウェアの配布形式は tar.gz (tarball)と rpm
- HiSeq X Ten データにも対応

ご質問やトラブルシュートのご相談は

イルミナテクニカルサポートにお問合せ可能

techsupport@illumina.com

ローカルLinux 上で使用する、FASTQ生成ツールの整理 (2014/11 現在)

bcl2fastq2 v2.15.0	NextSeq 500とHiSeq X Ten データをローカルLinuxサーバでFASTQに変換向け.
bcl2fastq v1.8.4	HiSeq データをローカルLinuxサーバでFASTQに変換向け. HCSによる圧縮データに対応している. ※ HiSeq V4データは圧縮されているため、旧CASAVA v1.8.2ではなく、こちらをお使い下さい. ※ MiSeqデータでの利用 はサポート外となりますがご使用頂けます. ※ 使い方は、2013/Oct/11 サポートウェビナーをご参考下さい。
CASAVA v1.8.2	・HiSeq, GAローカルLinuxサーバでFASTQに変換し、更にアライメントや変異コールもCASAVAで実施したい方向け. ・HCSによる圧縮データには未対応のため, 圧縮データ利用の場合はFASTQ変換まではbcl2fastq v1.8.4を用いる. ※ CASAVAの新規ご提供は終了致しました.

bcl2fastq2 インストール例

$ yum install -y bcl2fastq2-v2.15.0.4-linux-x86_64.rpm

弊社ホームページからダウンロードしたrpmパッケージ（ソフトウェア）

通常、 root権限が必要となります

ソフトウェアダウンロードとユーザーガイド；
http://support.illumina.com/downloads/bcl2fastq_conversion_software.html
UserGuide 中、 “Installing bcl2fastq2 ConversionSoftware”をご参考ください

bcl2fastq2 動作要件

メモリ 32GB 以上

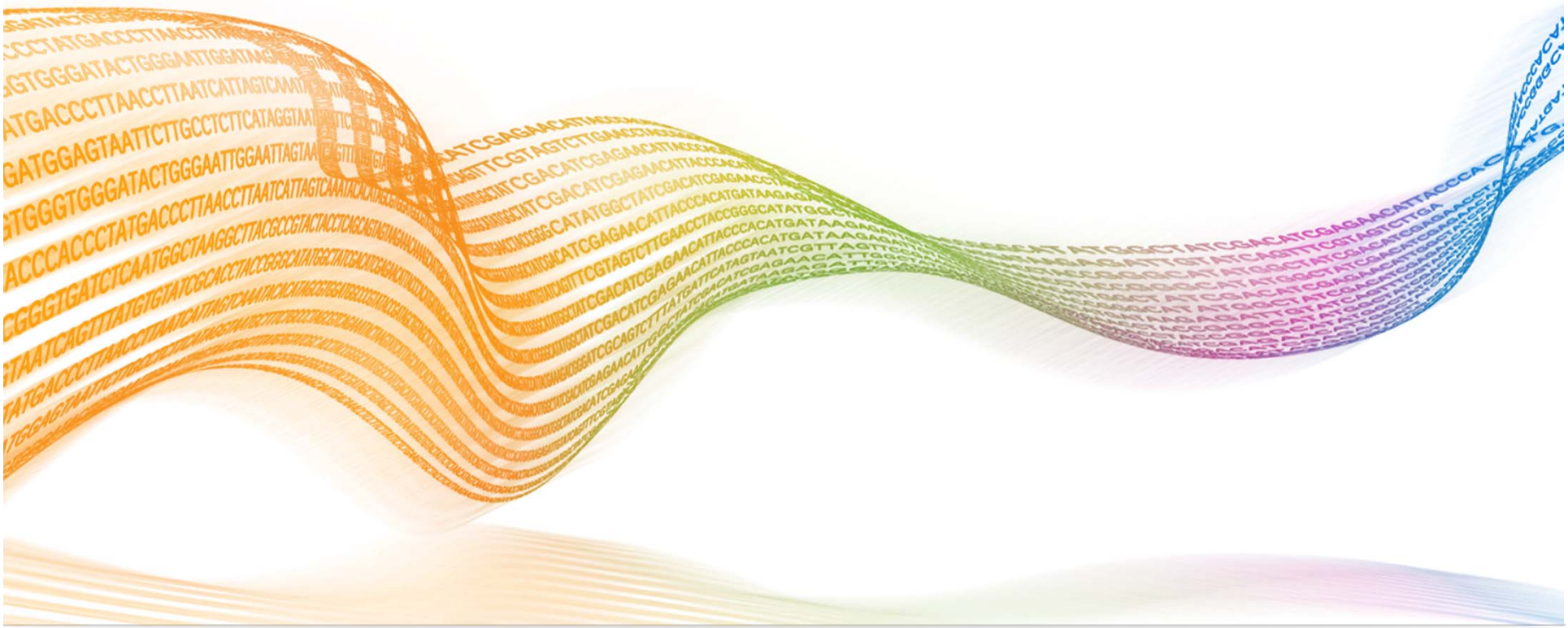
64bit CentOS か Red Hat Enterprise Linux
(テストは5でのみ実施)

インストールされている事が必要なライブラリ等；

zlib
librt
libpthread
gcc 4.1.2 (with c++)
boost 1.54 (with its dependencies)
cmake 2.8.9
zlib
librt
libpthread

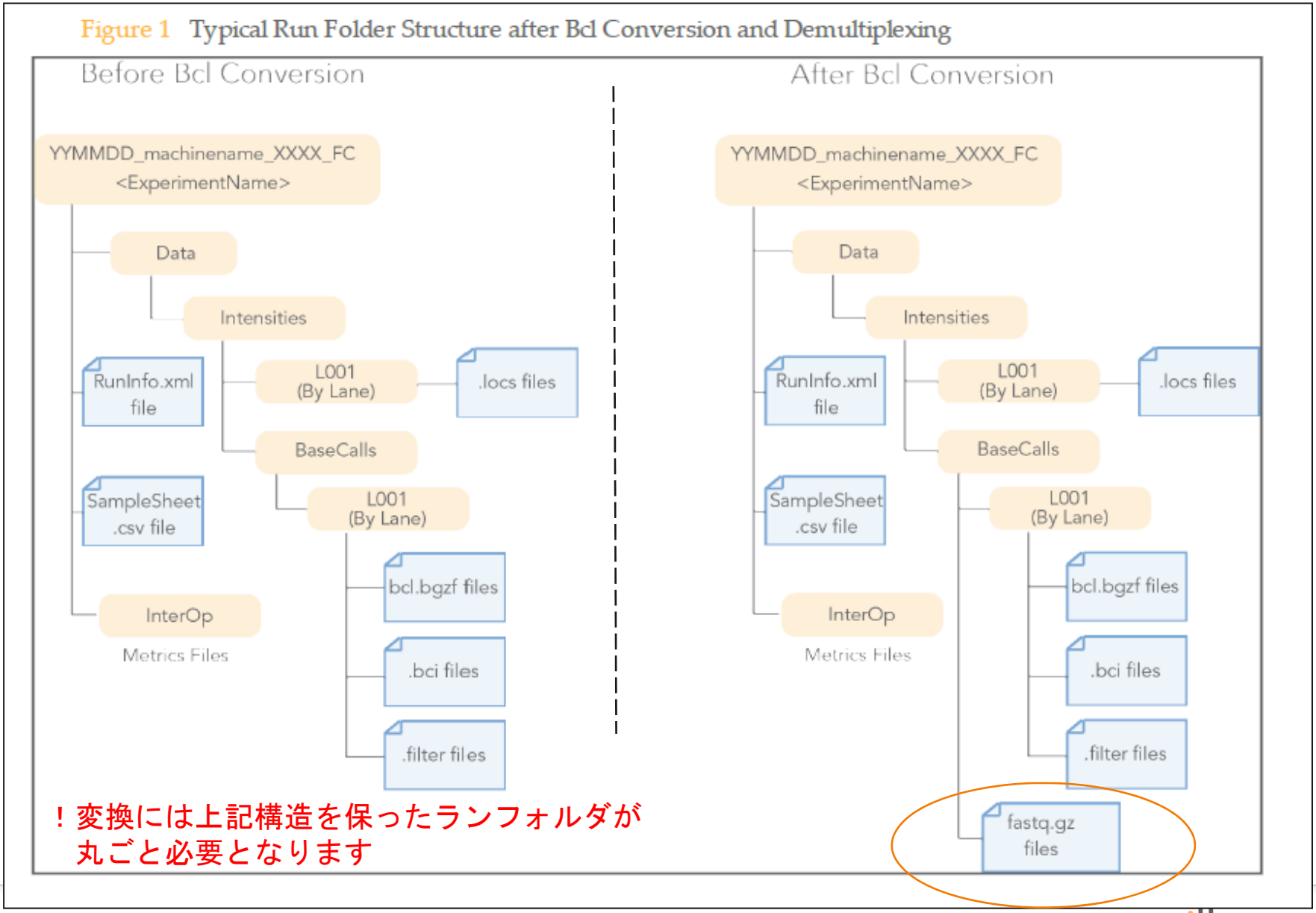
* bcl2fastq2 User Guide p.24 Appendix: Requirements

bcl2fastq2 実行前後のファイル構造と圧縮



illumina®

典型的なファイル構造 (通称：ランフォルダ)



illumina

データ圧縮機能のイメージ

※ 両圧縮とも、装置付属制御PCにて
圧縮実行されます

bclのZip圧縮および、幅を持ったQ Scoreの付与

30X human genome build size

110GB

デフォルト
= 圧縮なし

48GB

bgzip及び、非等間隔幅
を持たせた新 Qスコア
表現による圧縮

非等間隔

Qスコア	新Qスコア表現（非可逆圧縮）
2-9	6
10-19	15
20-24	22
25-29	27
30-34	33
35-39	37
≥40	40

bcl は >50%程度; BAMは ~30%程度のサイズ減少

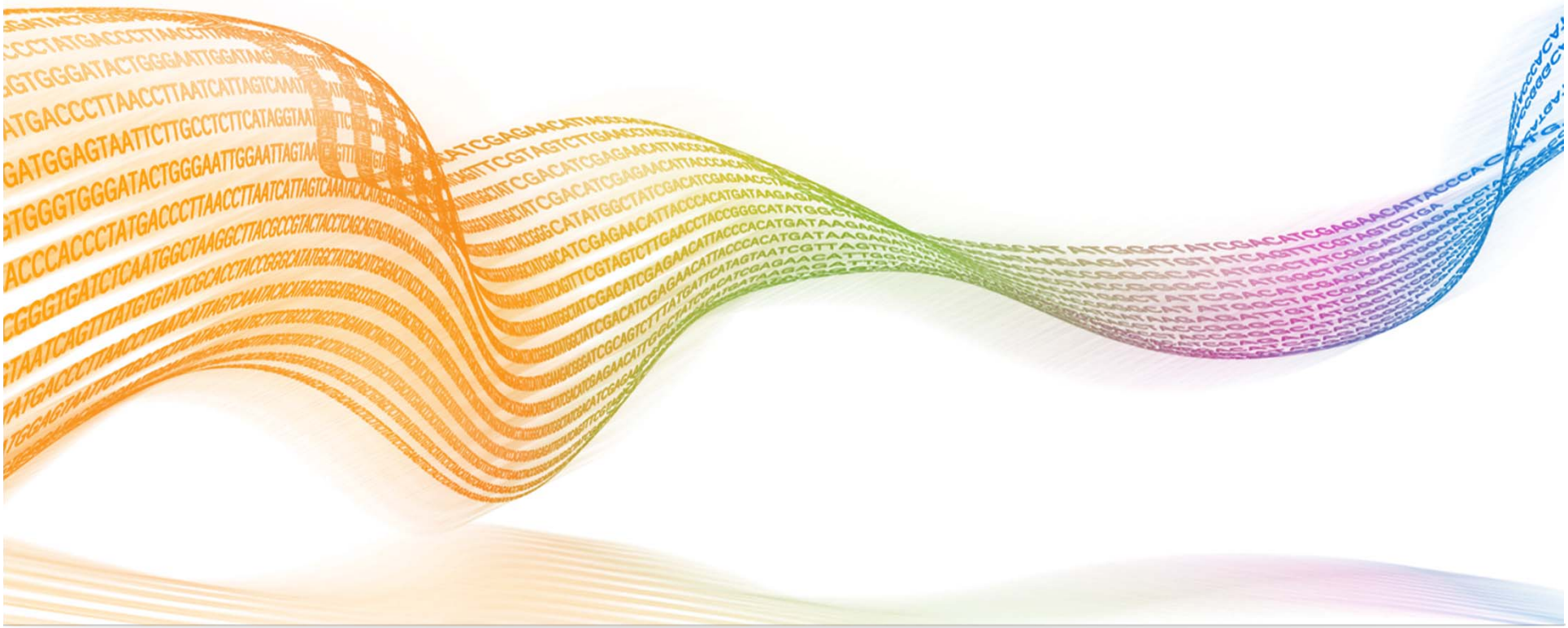
* http://www.illumina.com/Documents/products/whitepapers/whitepaper_datacompression.pdf

illumina

BGZF (Blocked GNU Zip Format)

圧縮形式のひとつ
gzip の拡張版
付加情報を持つため圧縮率はgzipより若干小さいが、プログラムからの情報アクセスが高速になる。
samtoolsで長い採用実績。
NextSeqの解析上は解凍せずに使用するが、一般的には通常のgzipクライアントで解凍できる。

bcl2fastq2　サンプルシートの準備



illumina®

NextSeq インデクス情報の与え方

ローカルサーバ利用の場合は、サンプルシートの作成が必要

サンプルシートをIEMで作成し、
bcl2fastq2 実行時に与える
(シーケンス時は利用せず).

bcl2fastq 2.0
ローカルサーバでFASTQ変換

BaseSpace
Prep タブ

BaseSpace

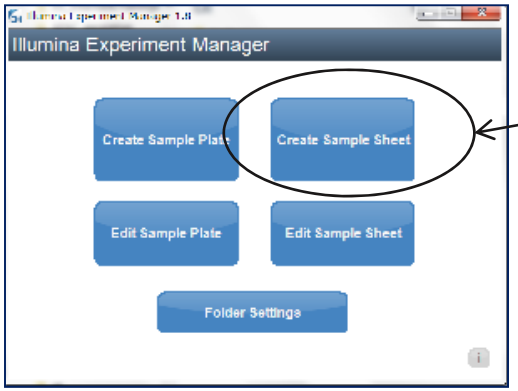
BaseSpaceクラウドや BaseSpaceOonsite 利用時は、
シーケンス開始前にBaseSpaceの　PrepTab機能を利用し
インデクス含めたサンプル管理情報を登録しておく。
シーケンス開始時にこれを指定する。

illumina

IEM を使ったサンプルシートの作成

IEM: Illumina Experimental Manager, Windows上で動作。
サンプルシート作成専用ウィザード

http://support.illumina.com/sequencing/sequencing_software/experiment_manager.html



1. Create Sample sheet 選択
2. NextSeqの画像を選択
3. NextSeq Fastq Onlyを選択
4. シーケンス長、インデクス等の入力を続け
5. 保存
6. ランフォルダ直下にSampleSheet.csvという名前でコピーしておく

illumina

サンプルシートの例

ヘッダ部

データ部

[Header]								
IEMFileVersion	4							
Investigator Name	Isabelle							
Experiment Name	HiSeq X ten run							
Date	3/14/2014							
Workflow	GenerateFASTQ							
Application	HiSeq FASTQ Only							
Assay	TruSeq HT							
Description	none							
Chemistry	Default							
[Reads]								
151								
151								
[Settings]								
ReverseComplement	0							
Adapter	AGATCGGAAGAGCACACGTCTGAACTCCAGTCA							
AdapterRead2	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT							
[Data]								
Lane	Sample_ID	Sample_Name	Sample_Plate	Sample_Well	I7_Index_ID	index	Sample_Project	Description
1	sample_ID1	test1			D701	ATTACTCG	HXten	
2	sample_ID2	test2			D712	AGCGATAG	Hxten	
3	sample_ID3	test3			D710	TCCGCGAA	Hxten	
4	sample_ID4	test4			D708	TAATGCGC	HXten	

*HiSeq X Tenの例（基本的に同じ書式）
*bcl2fastq2 User Guide p.21

[Header]	
IEMFileVersion	4
Date	2014/11/14
Workflow	GenerateFASTQ
Application	NextSeq FASTQ Only
Assay	TruSeq HT
Description	
Chemistry	Amplicon
[Reads]	
151	
151	
[Settings]	
Adapter	AGATCGGAAGAGCACACGTCTGAACTCCAGTCA
AdapterRead2	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

*NextSeq 500のヘッダ部の例、
下部にデータセクションが続く.
IEMウィザードにて作成

旧CASAVAをご存知の方向け；

CASAVAサンプルシートとの主な違い

- 旧来のCASAVAタイプのフォーマットでは無く、
  MSR,HASと同様、ヘッダのついたワークフロータイプのフォーマット

- Dual index はindex1列, Index2列 にそれぞれ別の列に記入。

- **名前はSampleSheet.csvである必要があり変更できない。**
  ->名前とファイルの置き場所でbcl2fastq2により自動認識される。
  -> ランフォルダ直下に配置する。
  -> --sample-sheetオプションはない。

・アダプタートリミング情報はサンプルシートに記入
（IEMで入力アシスト有り）

禁忌文字　(全システム共通)

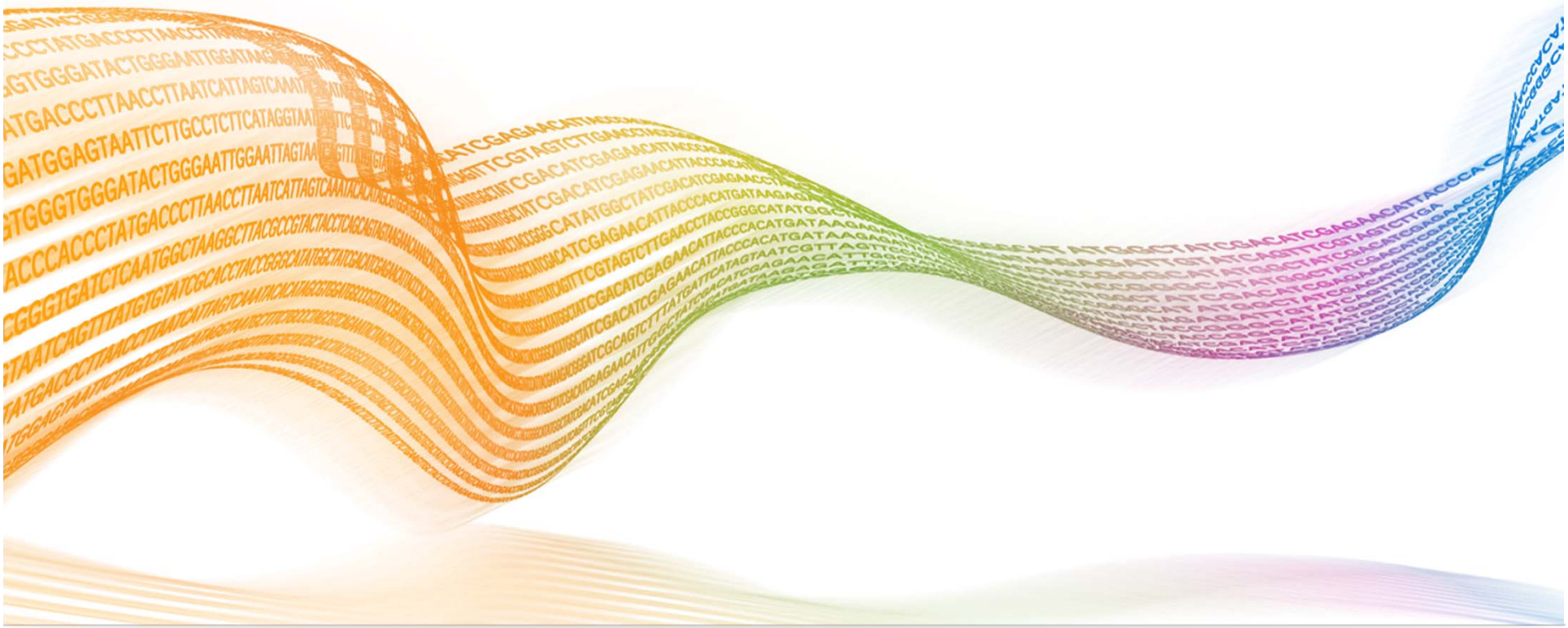
Illegal Characters

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

！これらの文字がサンプルシートに含まれますと、正しいエラーメッセージが
表示されないまま不正終了しますのでご注意ください。

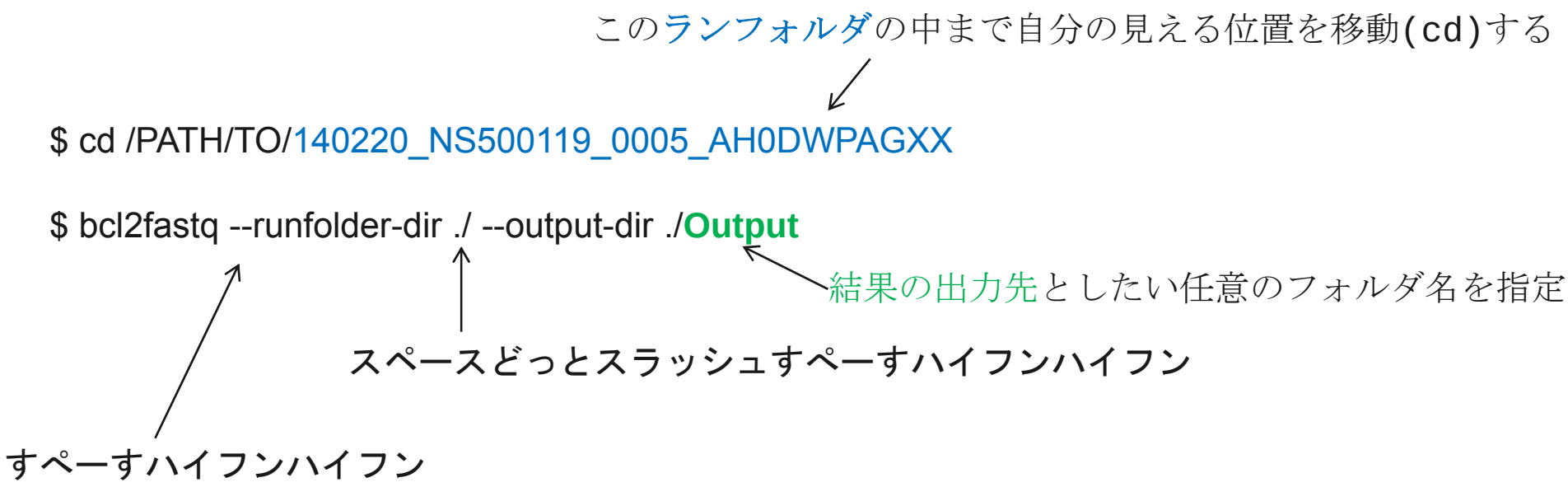
*bcl2fastq2 User Guide p.21

bcl2fastq2　実行



illumina®

bcl2fastq2 コマンドライン例



- PATH/TOの部分は、ご自分の環境に応じて変わりますので読み替えて下さい。
- 140220_NS500119_0005_AH0DWPAGXXはランフォルダ名で舞ラン毎に代わりますので
  こちらも都度読み替えて下さい。

illumina

主要な 指定解析パラメータ

オプション	内容
--barcode-mismatches	インデクス許容ミスマッチ デフォルト１ （インデクスごとに２まで指定可能）
--create-fastq-for-index-reads	インデクスFASTQを書き出す
--use-bases-mask	マスクする塩基を指定可能
--ignore-missing-bcls	欠損bclを無視する

他のパラメータにつきましては、(UserGuide p.15-17) をご参考ください。

実行時間 短縮 に重要な スレッド指定パラメータ（UserGuide p.15）

オプション	内容
-r, --loading-threads	BCLファイルロード用スレッド数
-d, --demultiplexing-threads	デマルチプレックス時利用スレッド数
-p, --processing-threads	デマルチプレックス後のデータ加工用スレッド数
-w, --writing-threads	FASTQ書き出し処理用スレッド数

実行時間とスレッド利用

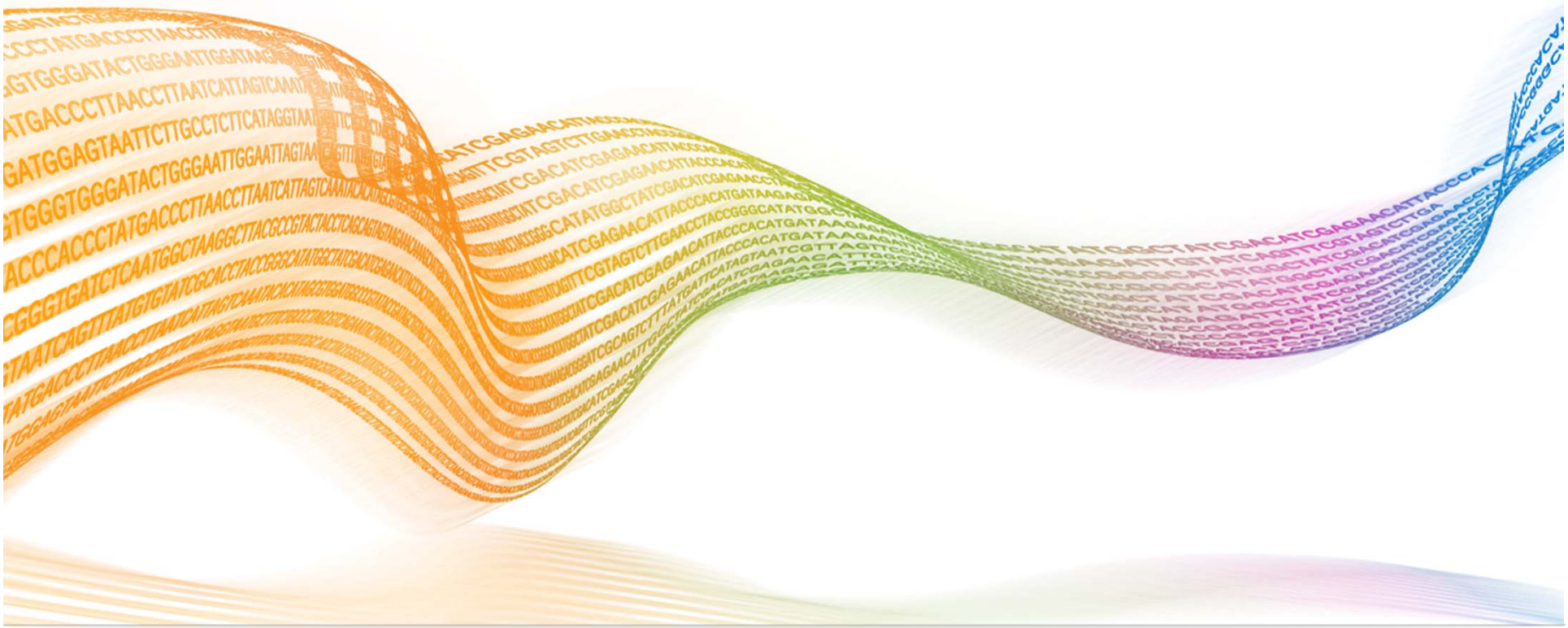
NextSeq 500のHOランについて、bcl2fastq2で FASTQ 変換を実施

マシンは 1CPU 16core 32GB mem

-r 2 -d 16 -p 16 -w 4	1:39:29	(CPU **1479**%)
no use (default)	14:34:41	(CPU **103**%)

※ベンチデータがございませんので、まずはご利用の環境でお試しください

bcl2fastq2　実行結果レポート



illumina®

デマルチプレックス結果簡易サマリの場所

Webブラウザで見られる簡易HTMLレポート；

ランフォルダ/Reports/html/ 配下

xmlファイル(HTMLレポートのデータファイル);

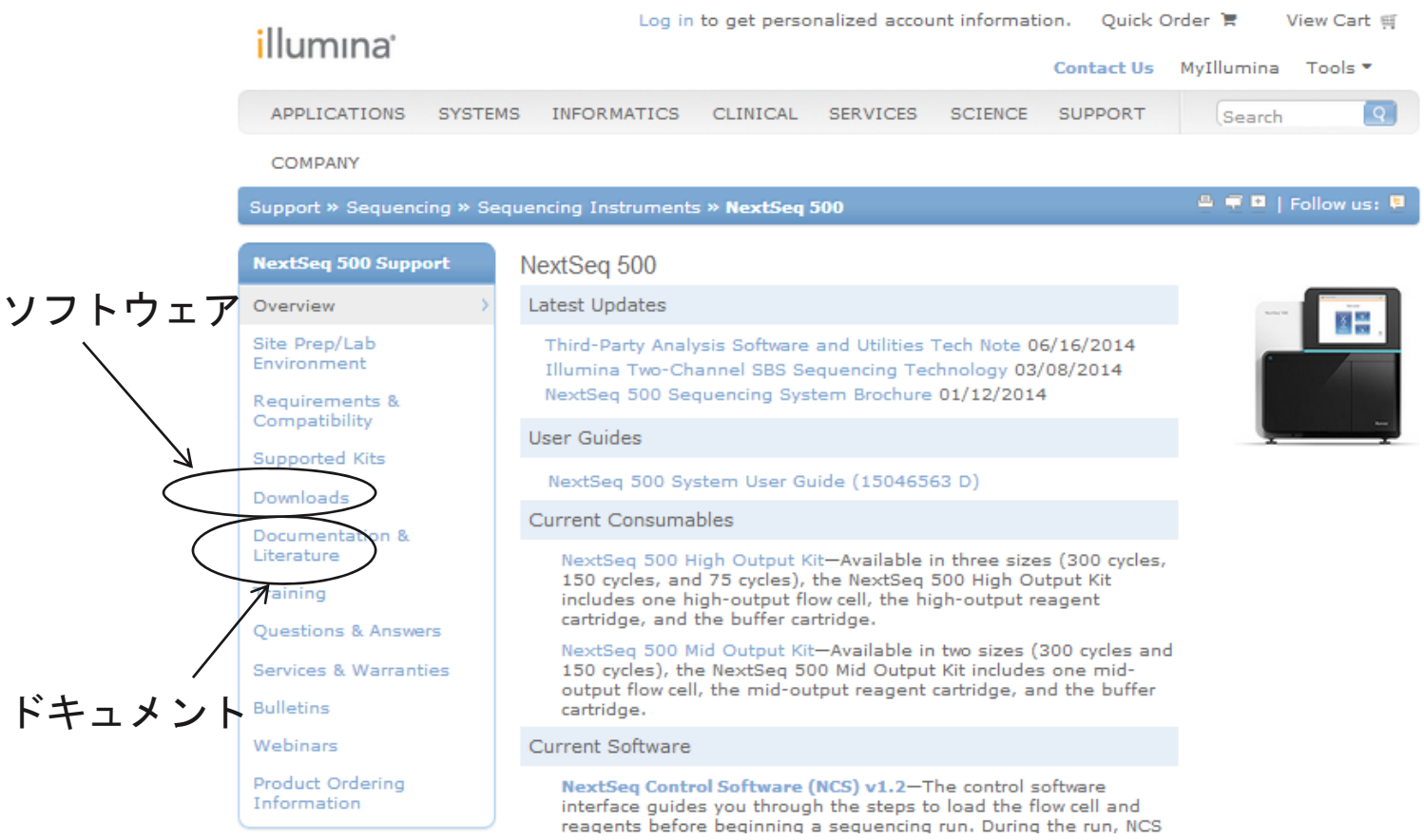
ランフォルダ/Stats/ConversionStats.xml
ランフォルダ/Stats/DemultiplexingStats.xml

NextSeq データと bcl2fastq2使用方法　まとめ

- NextSeqでは、bclファイルは圧縮されている。
- **Qscoreも圧縮される (通称：QScore binning)。変更不可。**
- レーン毎にひとまとまりの FASTQを作成するか(インデクス無の場合）、デマルチプレックスを実施し、レーン毎にサンプル毎のFASTQを生成。
- デマルチプレックスには、都度サンプルシートを作成しインデクス情報を記載する必要がある
- サンプルシートはIEMで作成することができる
- **CPU 2core以上でお持ちの場合、スレッドオプションが使え、HWリソースに応じて計算時間を短縮可能**
- そのほかサンプル振分け時のミスマッチ指定などオプション利用可能
- Linux を普段ご使用の方には難しくないレベルの使い勝手
- シーケンスやデマルチプレックスの結果がxml,htmlで出力される

NexSeq ポータルページ

- http://support.illumina.com/sequencing/sequencing_instruments/nextseq-500.ilmn



リソースページ

ソフトウェア

- http://support.illumina.com/sequencing/sequencing_instruments/nextseq-500/downloads.ilmn

ユーザガイド

- http://supportres.illumina.com/documents/documentation/software_documentation/bcl2fastq/bcl2fastq2-user-guide-15051736-b.pdf

NextSeqシステムについて

- http://supportres.illumina.com/documents/documentation/system_documentation/nextseq/nextseq-500-system-user-guide-15046563-d.pdf

Next is Now

Thank You!

illumina®