18th March 2021

Using Illumina Sequencing Technology to Explore The Bacterial Genome

Björn Espedido, PhD

Senior Field Application Scientist, Illumina

Zhiliang Chen, PhD

Senior Bioinformatics Sales Specialist, Illumina

illumina®

Agenda

SECTIONS

BaseSpace®
SEQUENCE HUB

1. Understanding the Illumina MiSeq system and library preparation
2. Introduction to BaseSpace Sequence Hub
3. Understanding bacterial whole genome alignment and variant calling

MiSeq and Library Preparation

1

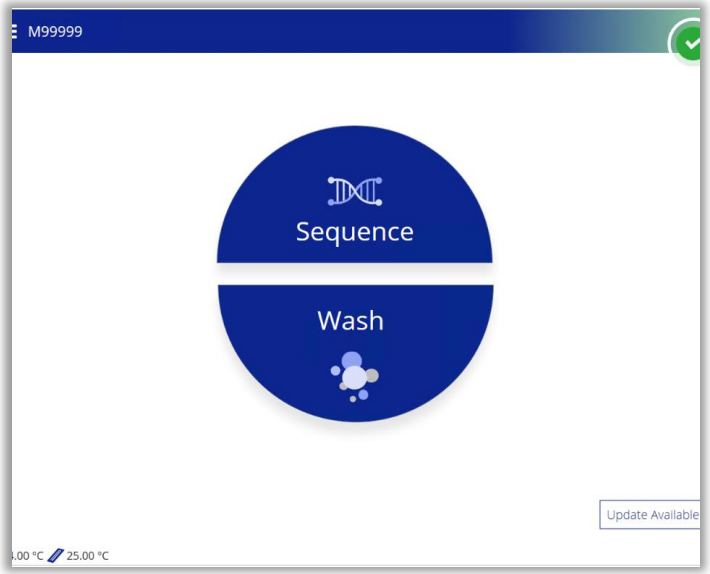
illumina®

MiSeq System Overview

MiSeq System Features

MiSeq Software

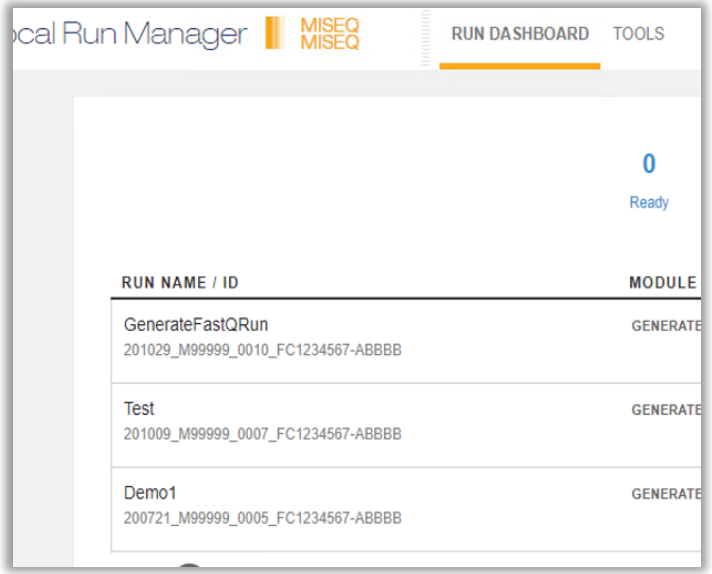
MiSeq Control Software (MCS)



Main functions

- System operation
- Performing sequencing runs
- Performing primary analysis

Local Run Manager (LRM)



Main functions

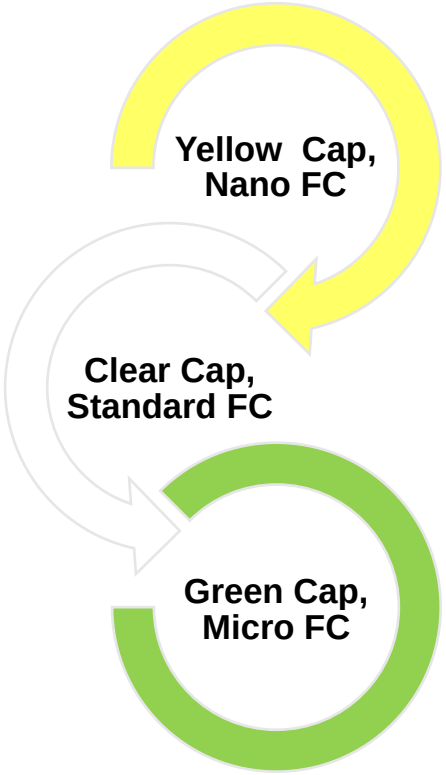
- Creating a run
- Run monitoring
- Performing Secondary analysis

MiSeq Sequencing Reagent Kit

Four components, all single-use

Box	Components	Storage
Box 1/2	① Reagent Cartridge	Freezer, -25°C to -15°C
	② HT1	
Box 2/2	③ Flow Cell	Refrigerator, 2°C to 8°C
	④ PR2 Buffer	

Flow Cell Types



Use the reagent cartridge associated with your flow cell type

MiSeq Reagent Kit Variations

Standard Flow Cells

MISEQ REAGENT KIT V2	
READ LENGTH	OUTPUT
1 × 36 bp	540-610 Mb
2 × 25 bp	750-850 Mb
2 × 150 bp	4.5-5.1 Gb
2 × 250 bp	7.5-8.5 Gb

MISEQ REAGENT KIT V3	
READ LENGTH	OUTPUT
2 × 75 bp	3.3-3.8 Gb
2 × 300 bp	13.2-15 Gb

Reads Passing Filter	V2	V3
Single reads	12-15 M	22-25 M
Paired- End Reads	24-30 M	44-50 M

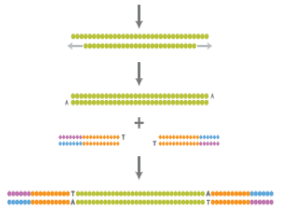
Low Output Applications

Flow cell	# of Reads	Read length	2 x 75 Output	2 x 150 Output	2 x 250 Output
Micro FC	4 M	Up to 2 x 150	600 Mb	1.2 Gb	----
Nano FC	1 M	Up to 2 x 250	150 Mb	300 Mb	500 Mb

Library Prep

Illumina Sequencing Workflow

1. **Library Preparation**



2. **Cluster Generation**

3. **Sequencing**

4. **Data Analysis**

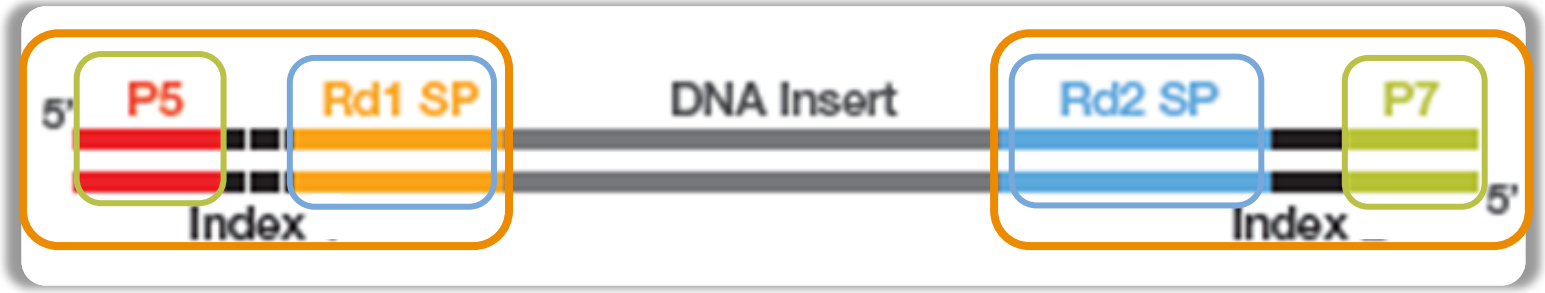
Planning Your Prep and Run

- Before proceeding to library prep and creating your run, you will need to plan your run by defining the following parameters of your sequencing run:

Establish an index scheme	• How many samples do I want to multiplex?
Identify the pooling strategy	• Which libraries will I sequence together?
Define sequencing read length	• What are my analysis requirements?

Library Prep is Critical for Successful Sequencing

The aim of library prep is to obtain nucleic acid fragments with adapters attached on both ends



Dual Index Library shown

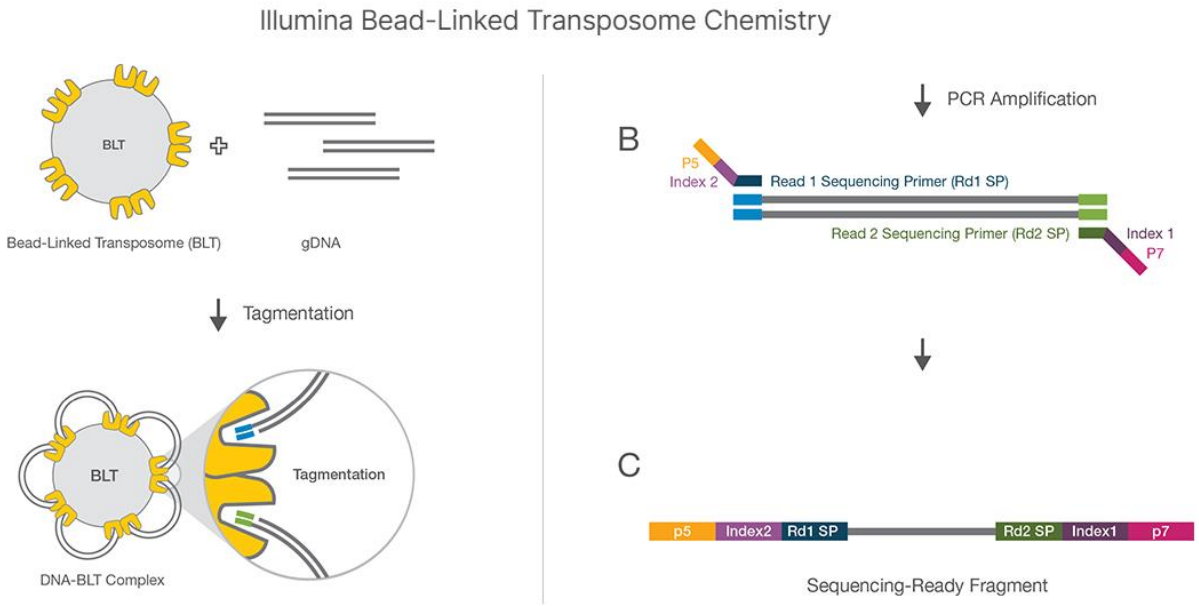
P5 and P7 sequences are complimentary to oligos bound to flow cell surface and *required* for any library

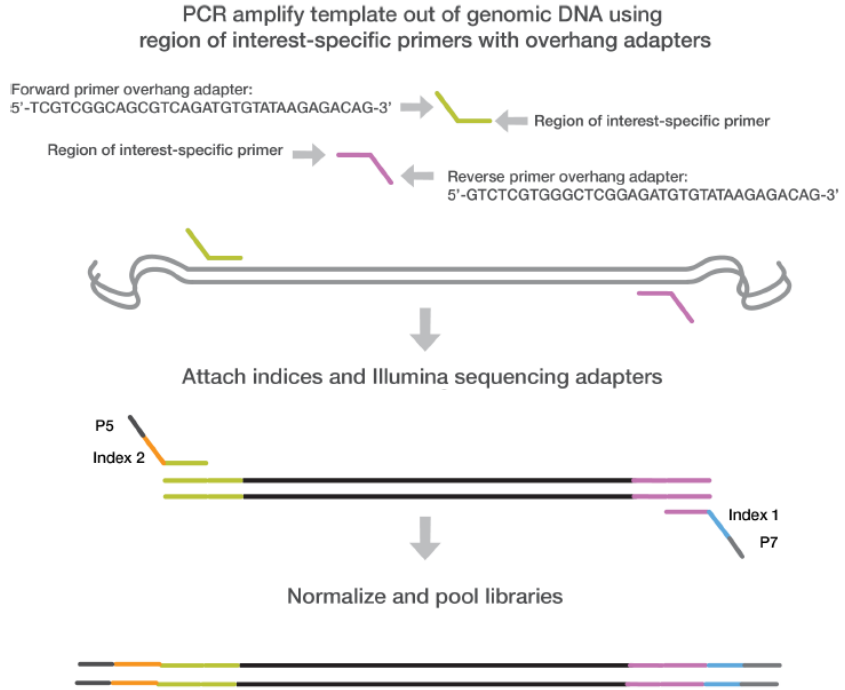
Indexes are used to tag individual samples to allow pooling of multiple libraries

Rd1 and Rd2 sequencing primers regions are used to initiate sequencing

Library Prep

Illumina DNA Prep, Tagmentation and DNA Indexes



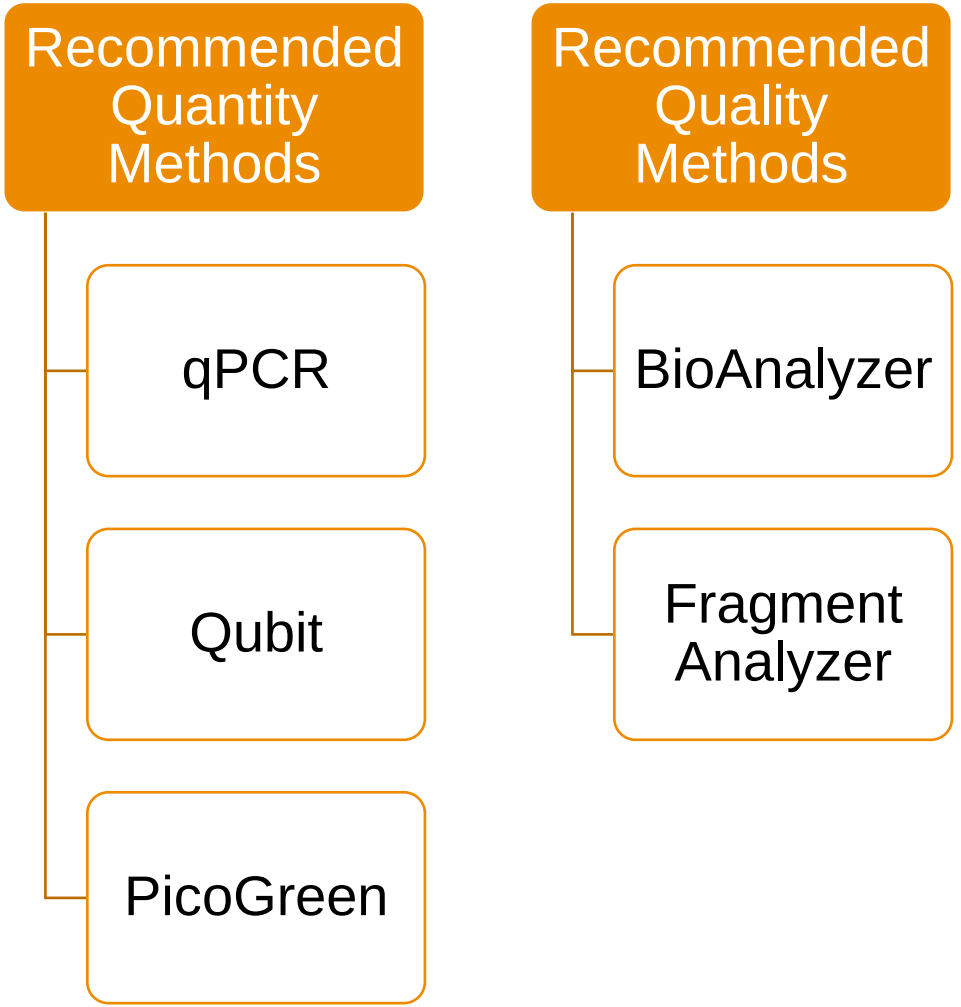


For large amplicons & whole genome sequencing

Amplicon sequencing using custom locus-specific primers with Illumina DNA indexes

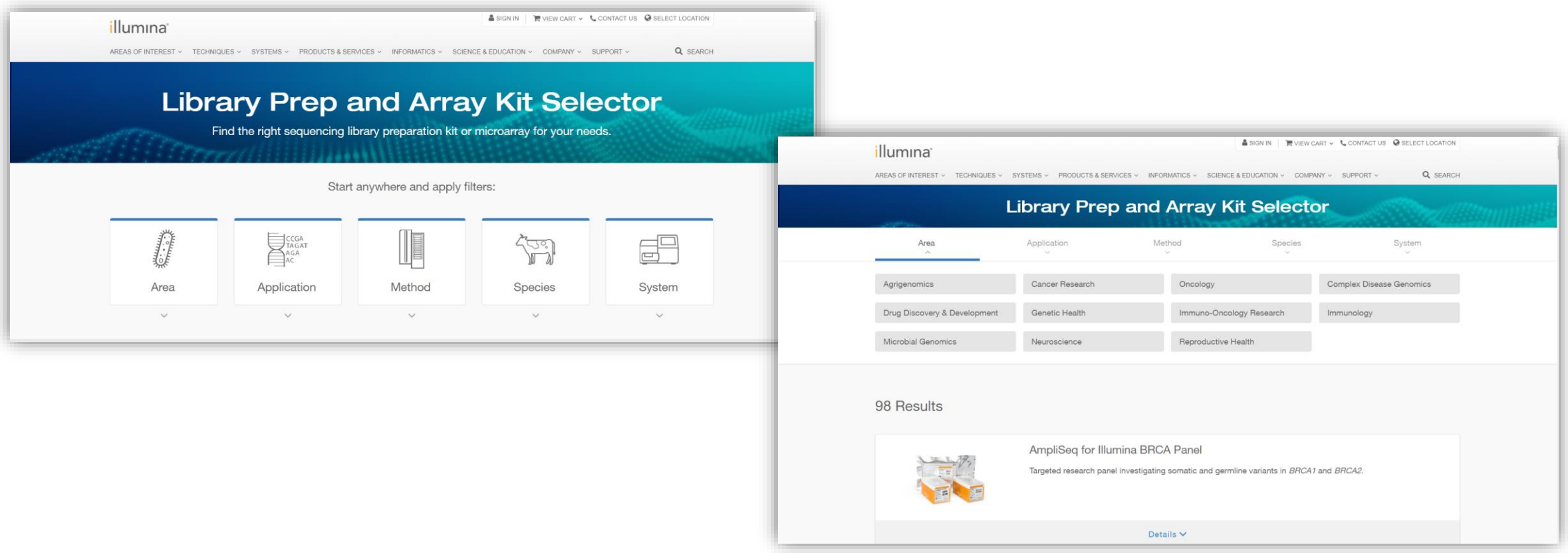
Library Quality and Quantity Methods

- How are you planning to assess library quality and quantity?
  - Accurate quantification of quality libraries is critical for optimal cluster density and data quality



Not Sure Which Kit to Use?

Use the Sequencing Sample Preparation Kit Selector



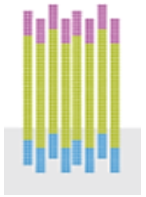
http://www.illumina.com/library-prep-array-kit-selector.html

MiSeq Sequencing Technology

Illumina Sequencing Workflow

1 Library Preparation

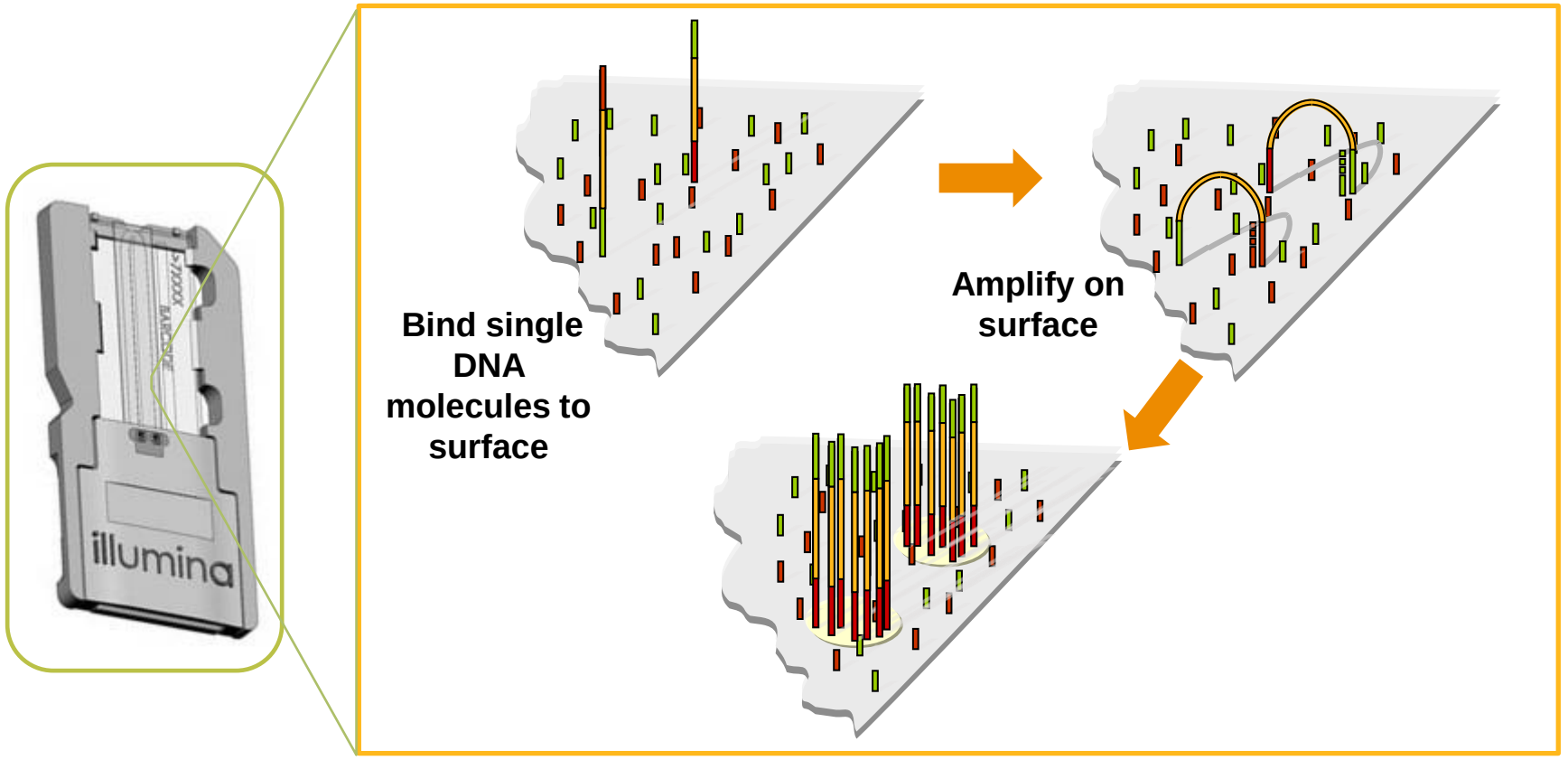
2 Cluster Generation



3 Sequencing

4 Data Analysis

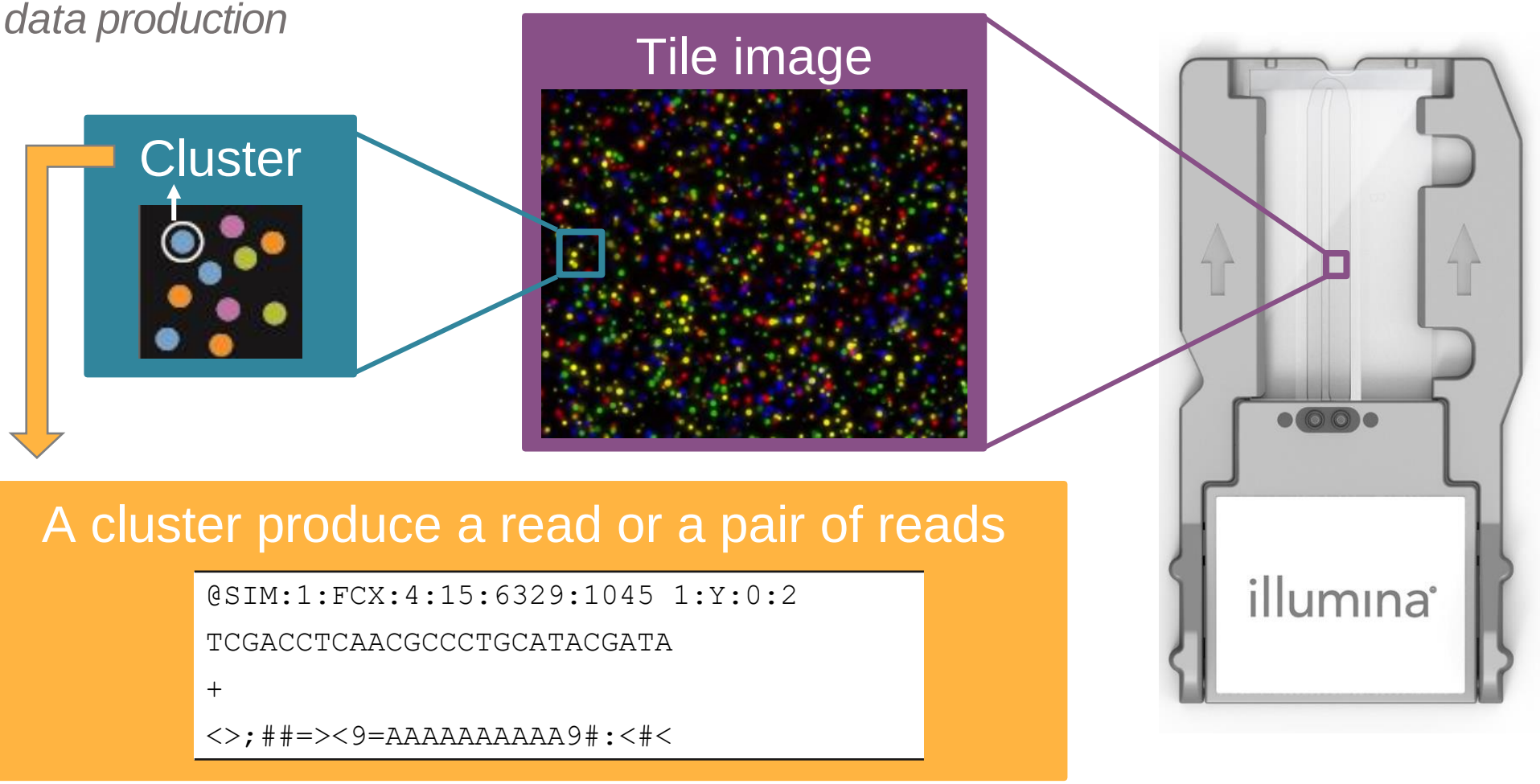
Cluster Generation Overview



If the internet is available, click here to view a video:

Data Yield Increases as Number of Usable Cluster Increases

Cluster is unit of data production



A cluster produce a read or a pair of reads

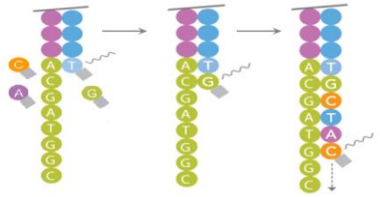
```
@SIM:1:FCX:4:15:6329:1045 1:Y:0:2
TCGACCTCAACGCCCTGCATACGATA
+
<>;##=><9=AAAAAAAAA9#:<#<
```

Illumina Sequencing Workflow

1 Library Preparation

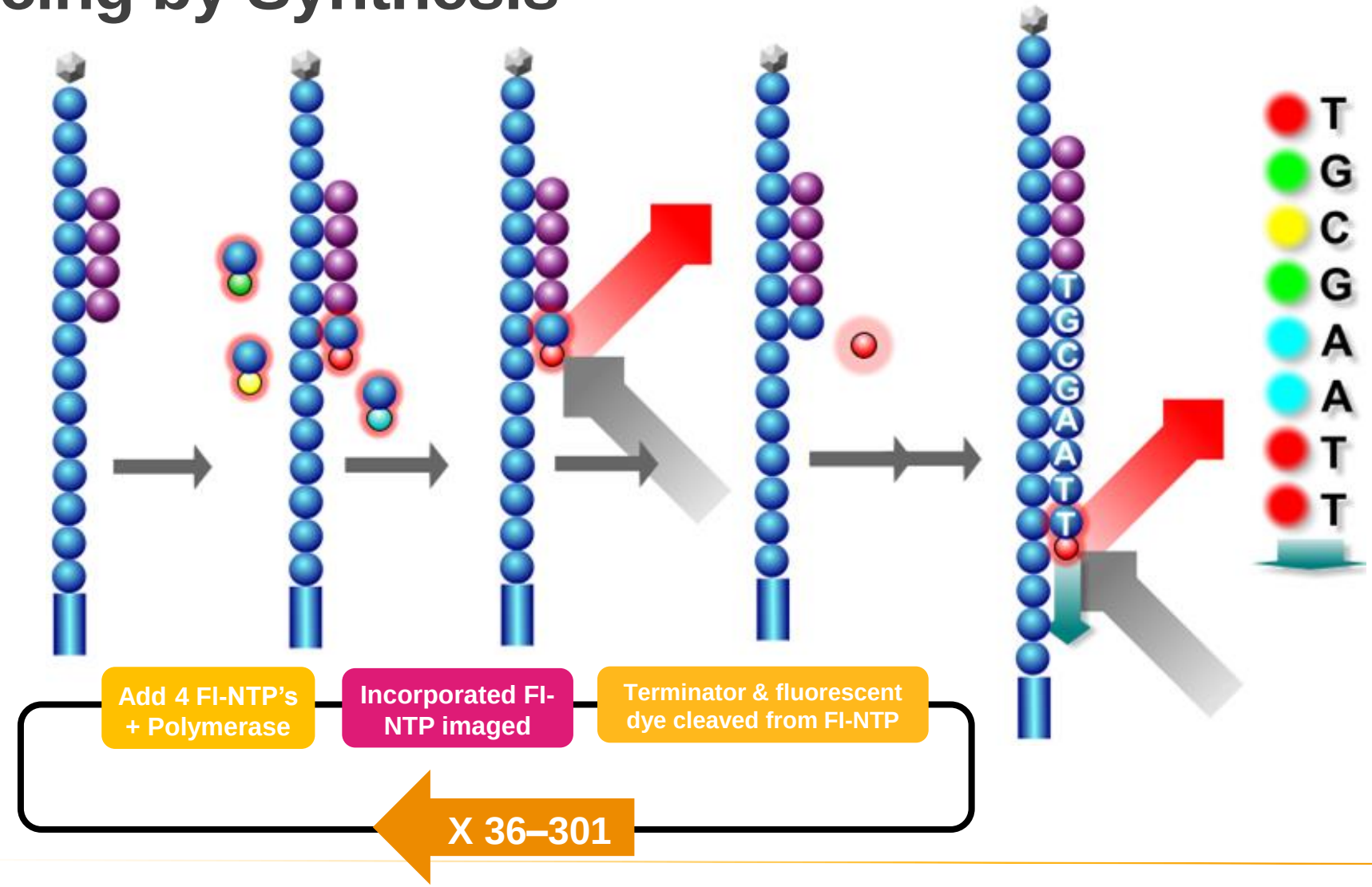
2 Cluster Generation

3 Sequencing



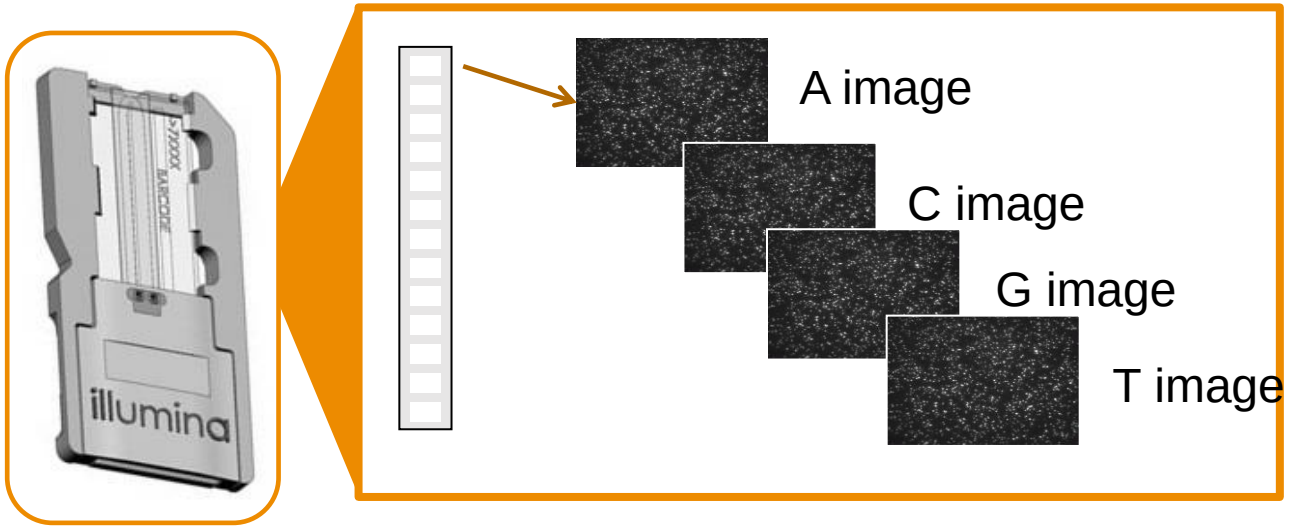
4 Data Analysis

Sequencing by Synthesis



Imaging on the MiSeq

- Clusters are imaged using LEDs, cameras and filter combinations specific for each fluorescently labeled nucleotide



- After imaging is complete for one section (tile), the flow cell is moved to the next tile and the process is repeated

Illumina Sequencing Workflow

1. **Library Preparation**

2. **Cluster Generation**

3. **Sequencing**

4. **Data Analysis**

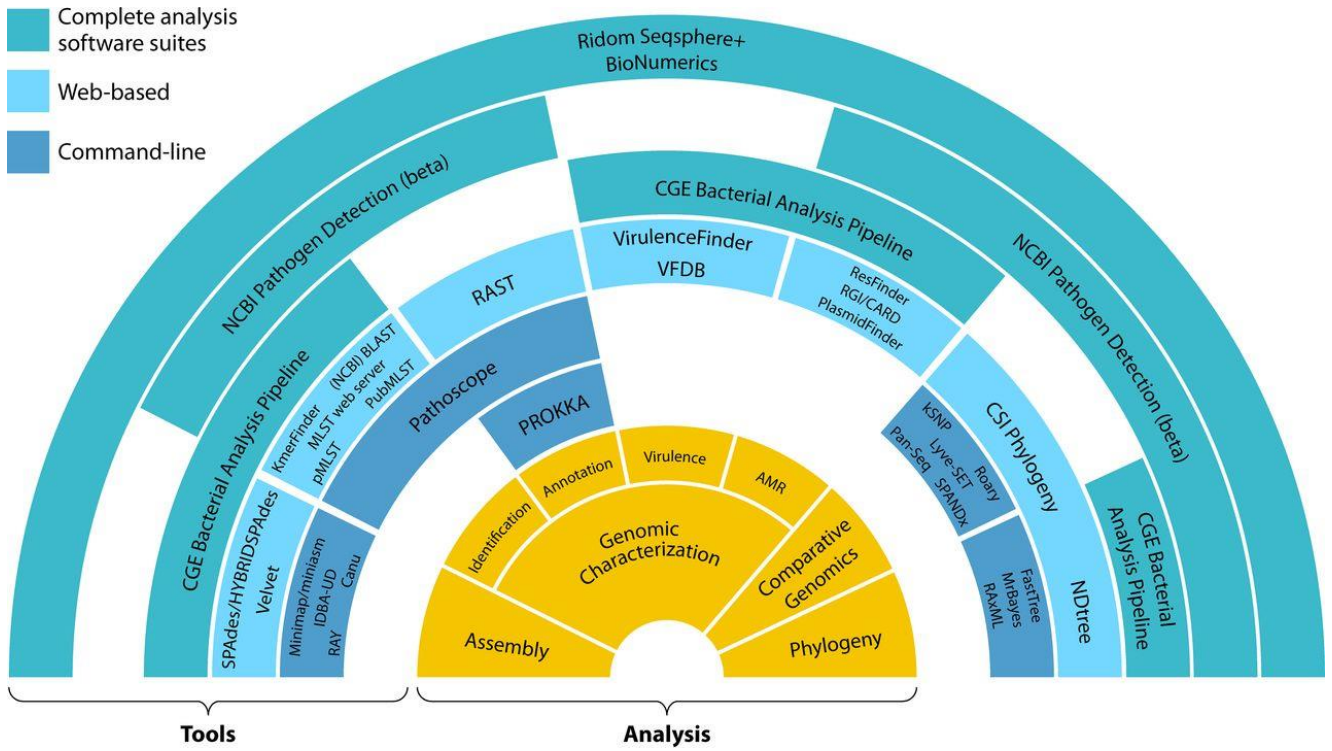
SUBTITLE

Introduction to BaseSpace Sequence Hub

illumina®

Analysis Tools

Diverse and growing set of tools allow for easier analysis of WGS data and new insights

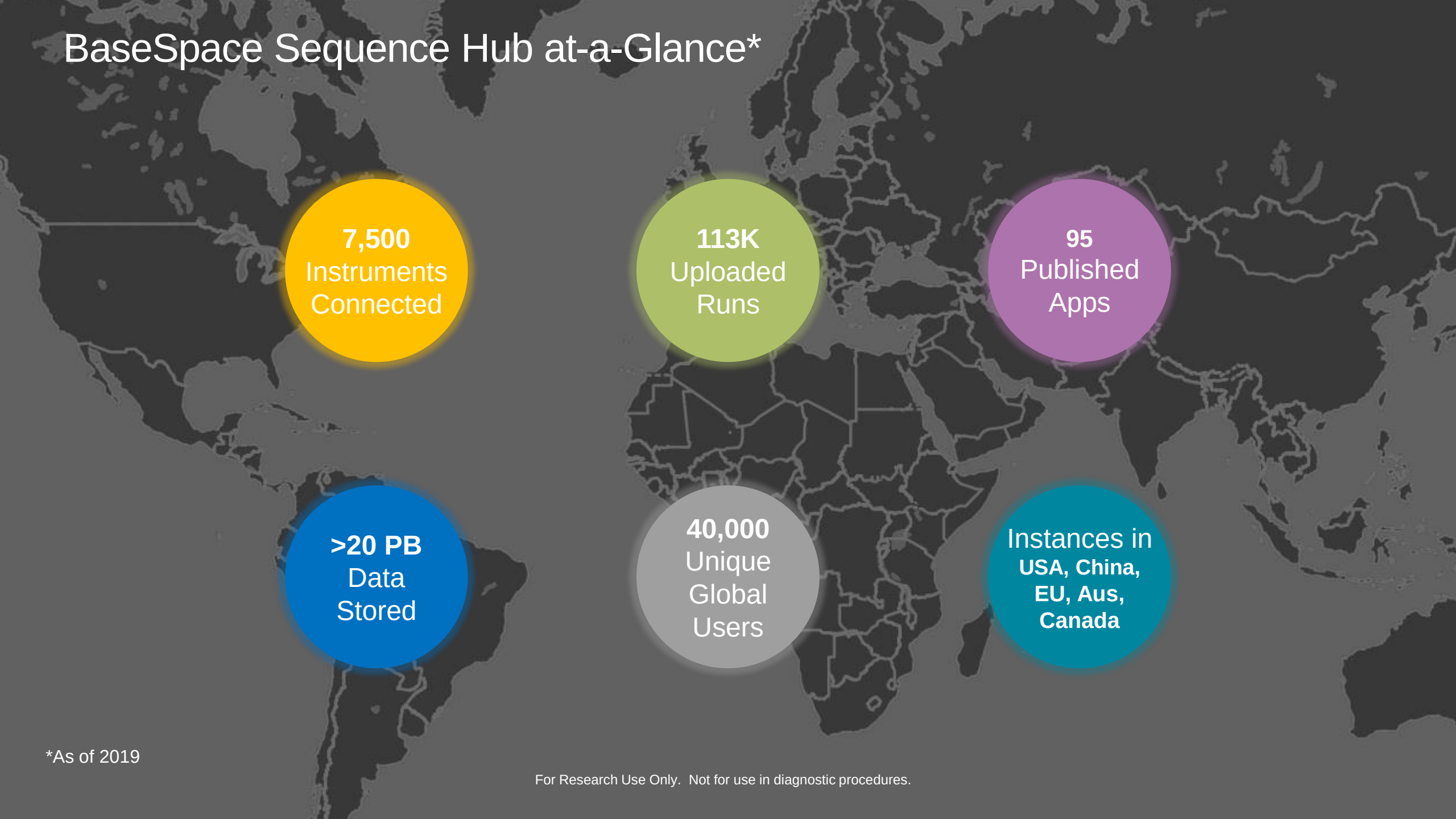


Clinical Microbiology Reviews Aug 2017, 30 (4) 1015-1063; DOI: 10.1128/CMR.00016-17

Vision

The destination for genomic investigators to **manage** and **analyze** genomic data, collaborate, and drive towards biological insights

BaseSpace Sequence Hub at-a-Glance*



*As of 2019

Accessing BaseSpace Sequence Hub

- US Instance
  - AWS Virginia
  - More compute power
  - https://basespace.illumine.com
- AU Instance
  - AWS Sydney
  - Data Stores within Australia
  - https://aps2.sh.basespace.illumine.com

- 2 instances share the same set of credentials
  - Data cannot be shared/transferred across regions

BaseSpace™ Sequence Hub

Enables anyone to analyze their NGS data

Plug and play with tight instrument integration

Easy **sharing and collaboration** worldwide

Simple **push-button analysis** with public and private analysis tools

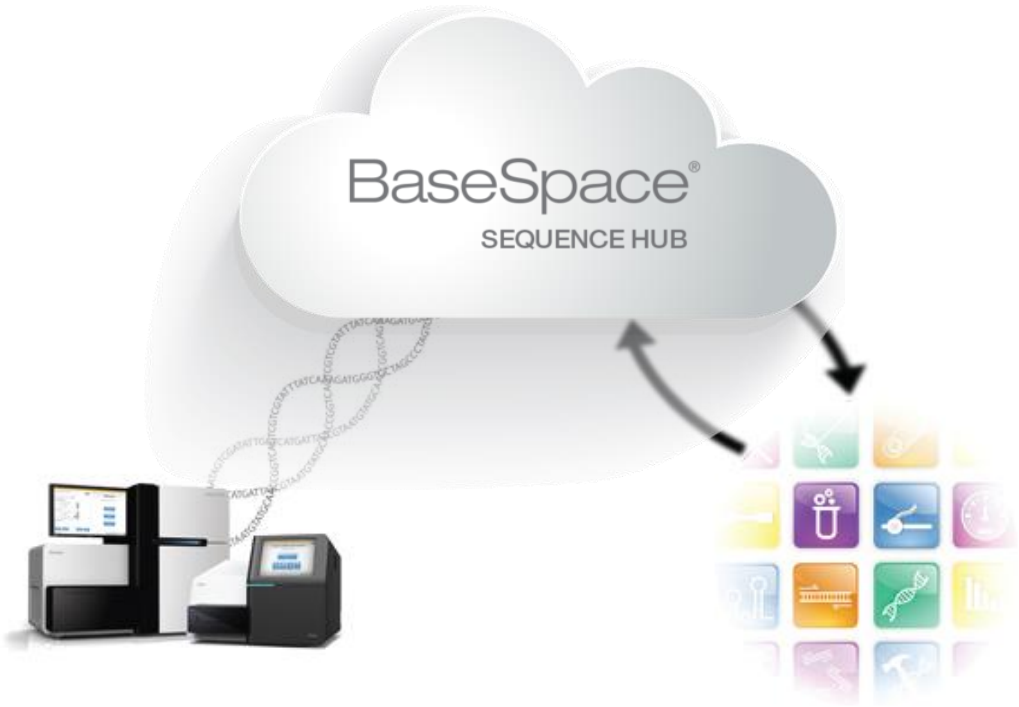
Advanced **automation and integration**

Comprehensive **security and privacy** controls

BaseSpace™ Sequence Hub

Genomics computing environment for NGS data

Monitor run quality, share data



Share and Collaborate

Upload directly from instrument

Use BaseSpace apps to interact with your data

Broad Range of Sequence Hub Applications for Secondary Analysis

BaseSpace SEQUENCE HUB

16s Metagenomics App

MetaPhlan Huttenhower Lab, HSPH

Pedant SequenceAnalyzer Biomax Informatics AG

Prokka Genome Annotation Basepace Labs

De novo assembly

DRAGEN Germline

Generate FASTQ

SPAdes Genome Assembler

DNASTAR® SeqMan Ngen Assembler

Velvet de novo Assembly Basespace Labs

NCBI SRA submission

Broad IGV genome visualization

Kraken Metagenomics App

DeepChek®-HBV ABL S.A.

DRAGEN Metagenomics Pipeline

QIIME Preprocessing QIIME Development Team

SRST2 BaseSpace Labs

Genius Metagenomics: Know Now, CosmosID

E. coli Stereotyping GoSeqIt

SEAR: Antibiotic Resistance University of Cambridge

Chunlab Trubac ID

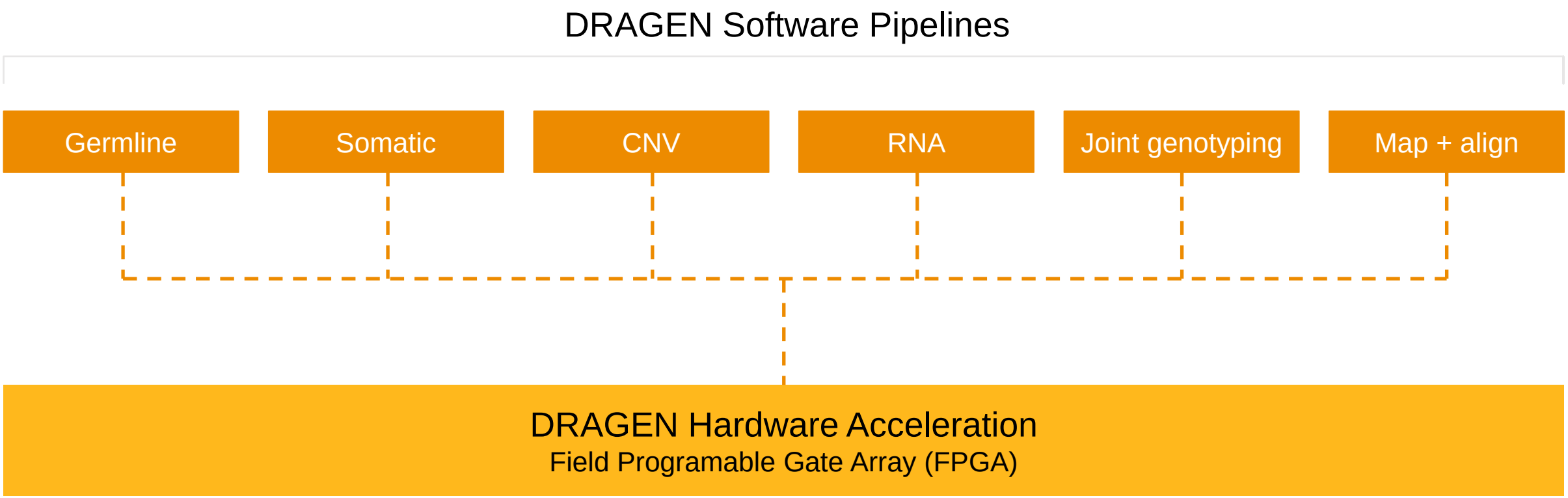
bioMerieux EPISEQ

IDbyDNA Infectious Disease Analysis Platform

DRAGEN COVID Lineage

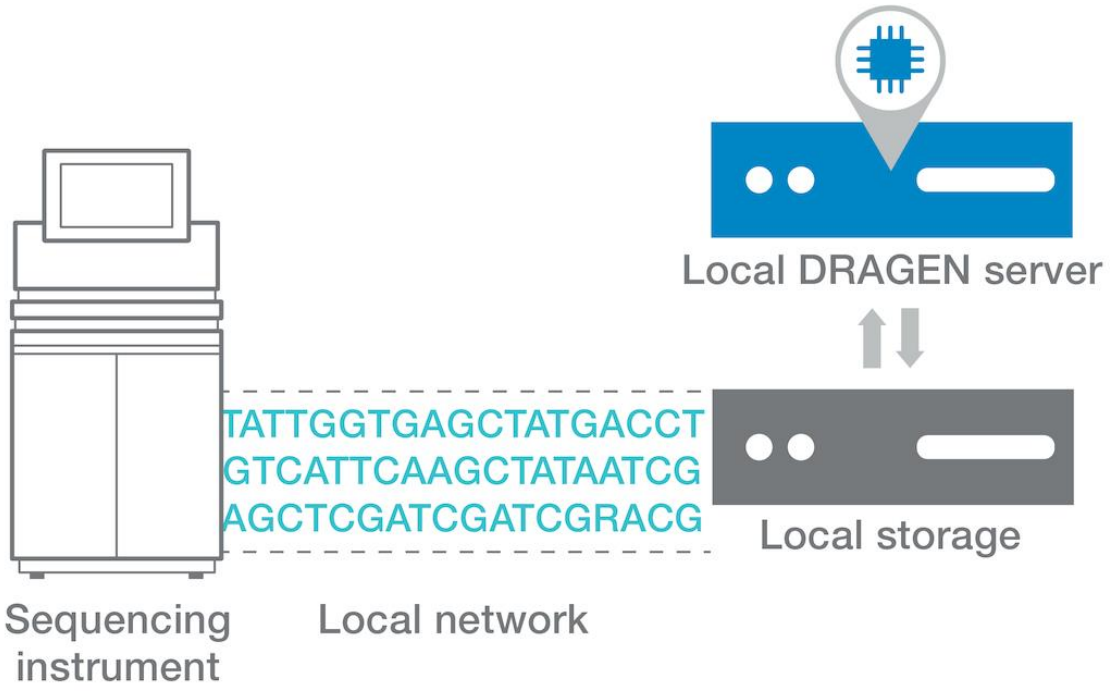
DRAGEN™ is Hardware-Accelerated Secondary Analysis

Dynamic Read Analysis for GENomics

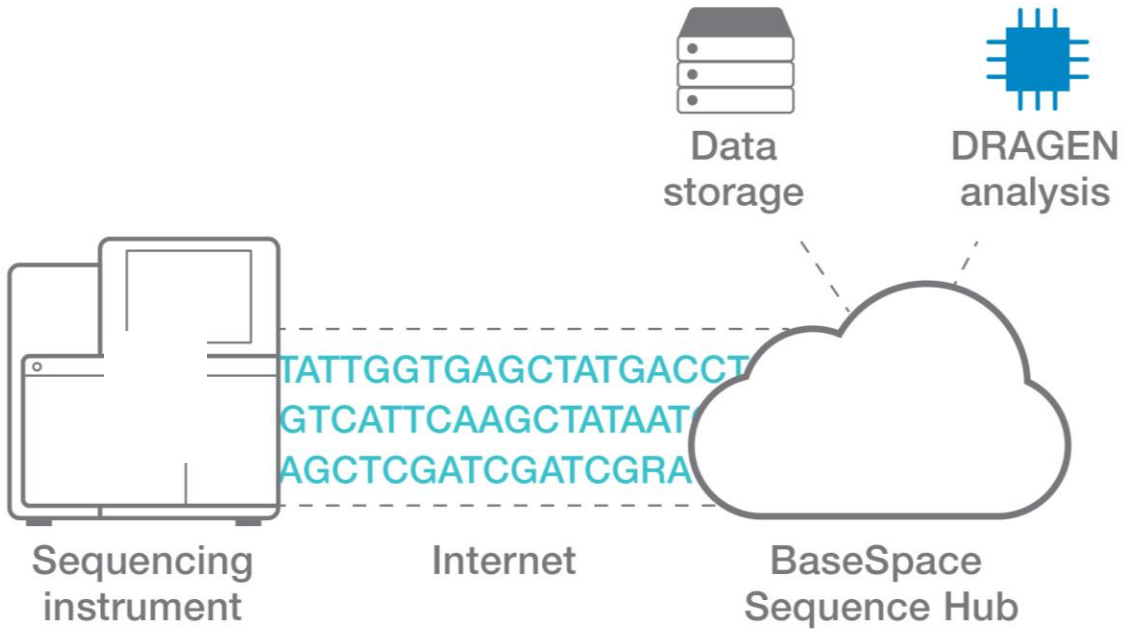


Flexible Data Analysis

On-premise



BaseSpace™ Sequence Hub



DRAGEN on BaseSpace Sequence Hub

Accurate, rapid secondary analysis in an easy-to-use, cloud-based environment

Accurate, Fast Analysis

Simple Workflow

Low-cost, Scalable Platform

Secure, Compliant Environment

Available Pipelines

- DRAGEN Germline Pipeline
- DRAGEN Somatic Pipeline
- DRAGEN Enrichment Pipeline
- DRAGEN RNA Pipeline
- DRAGEN Joint Genotyping Pipeline
- DRAGEN Methylation Pipeline
- DRAGEN Reference Builder

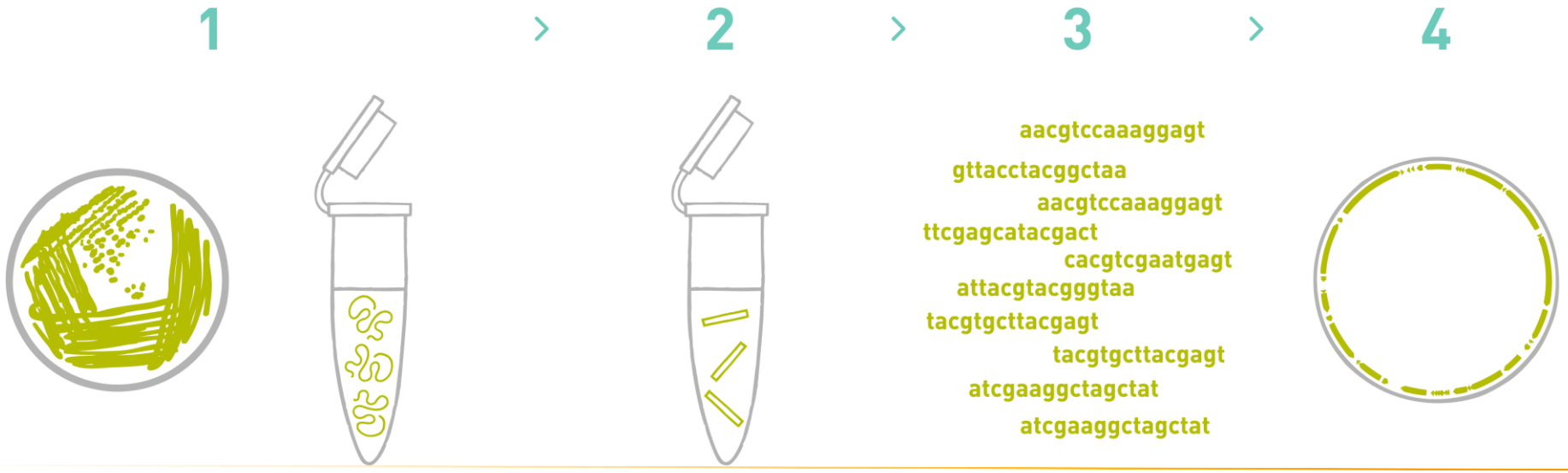
Bacterial Whole Genome Alignment and Variant Calling

3

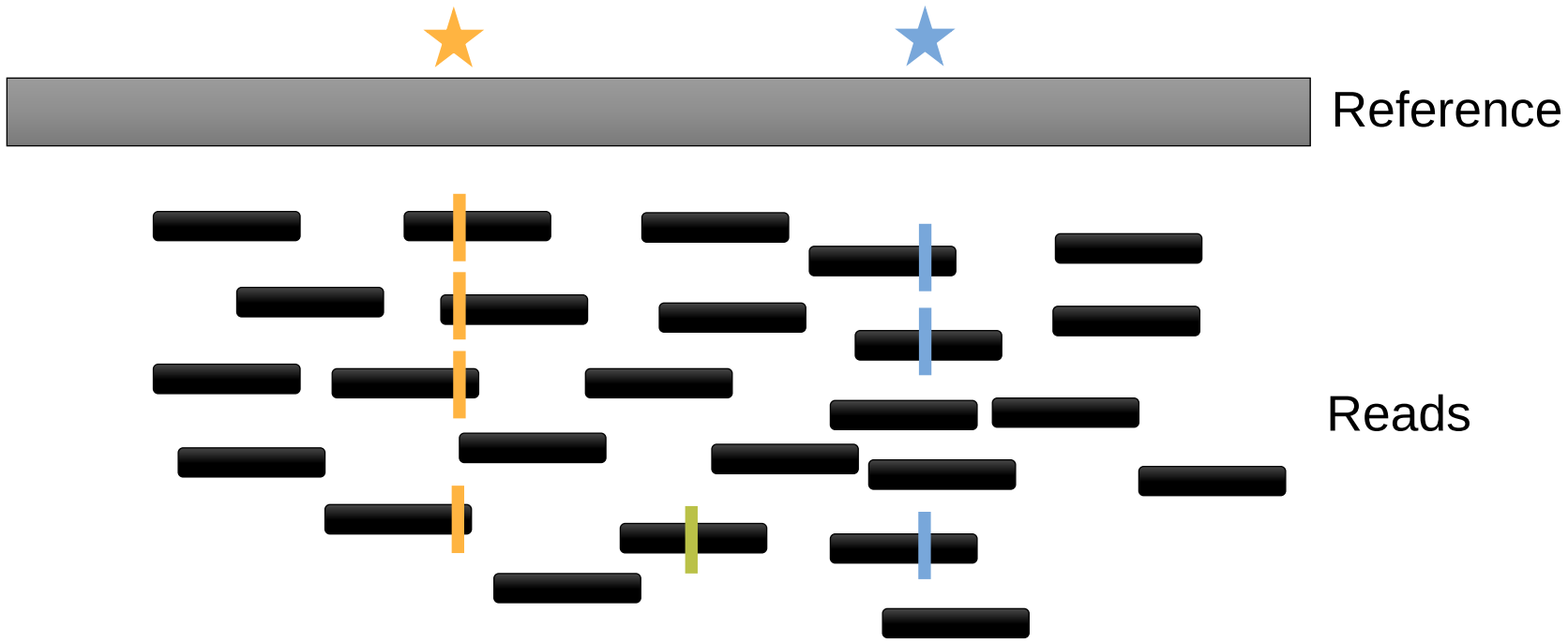
illumina®

Bacterial Whole genome resequencing

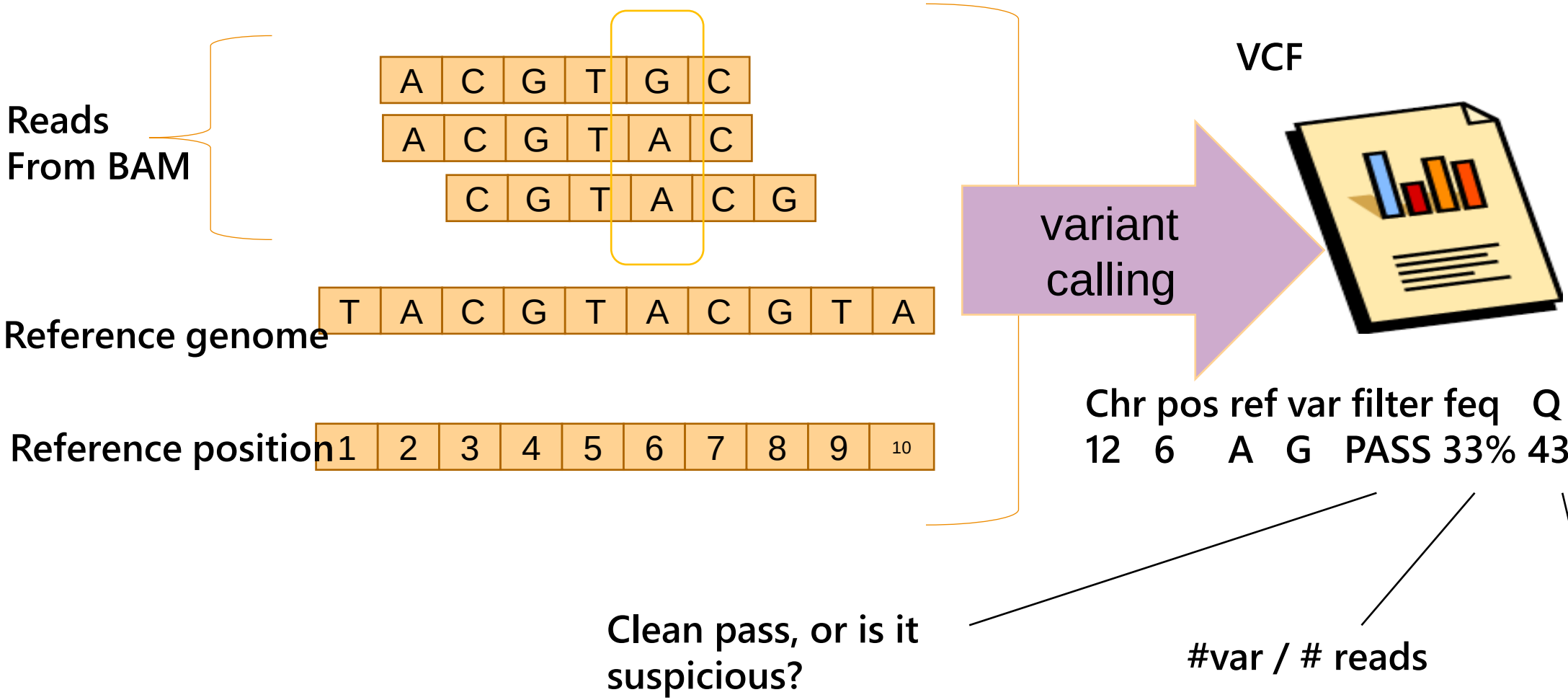
- Microbial whole-genome resequencing involves sequencing the entire genome of a bacteria, virus, or other microbe, and comparing the sequence to that of a known reference.
- The Germline (resequencing) workflow compares the DNA sequence in the samples against a reference genome and identifies any variants (SNPs or indels) relative to the reference sequence.



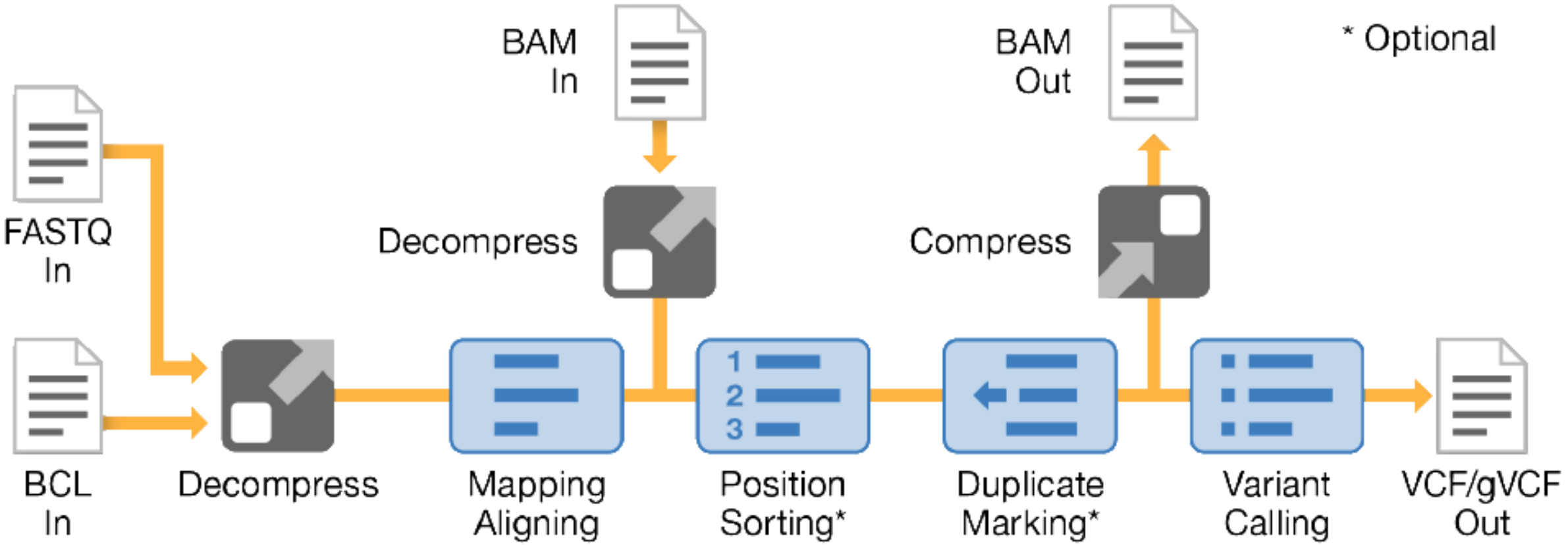
Whole Genome Resequencing



What is variant calling?

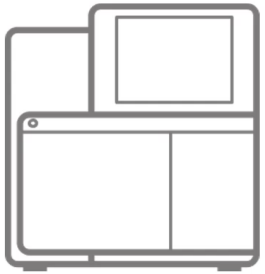


DRAGEN Germline Pipeline



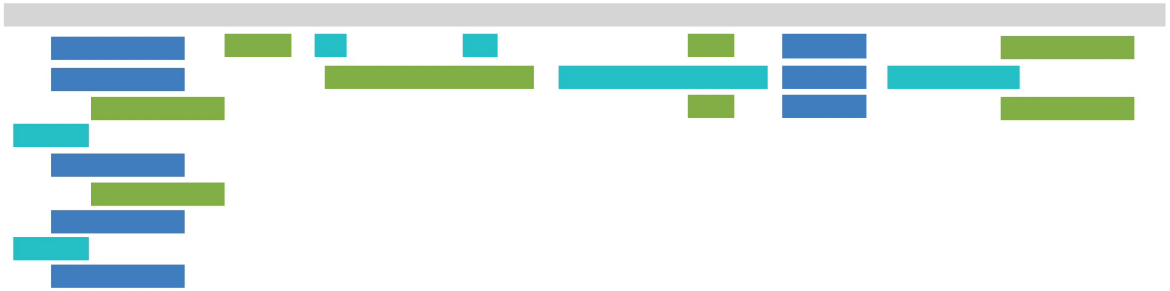
Mapping & Aligning—Germline Pipeline Example

Mapping Aligning | Position Sorting* | Duplicate Marking* | Variant Calling



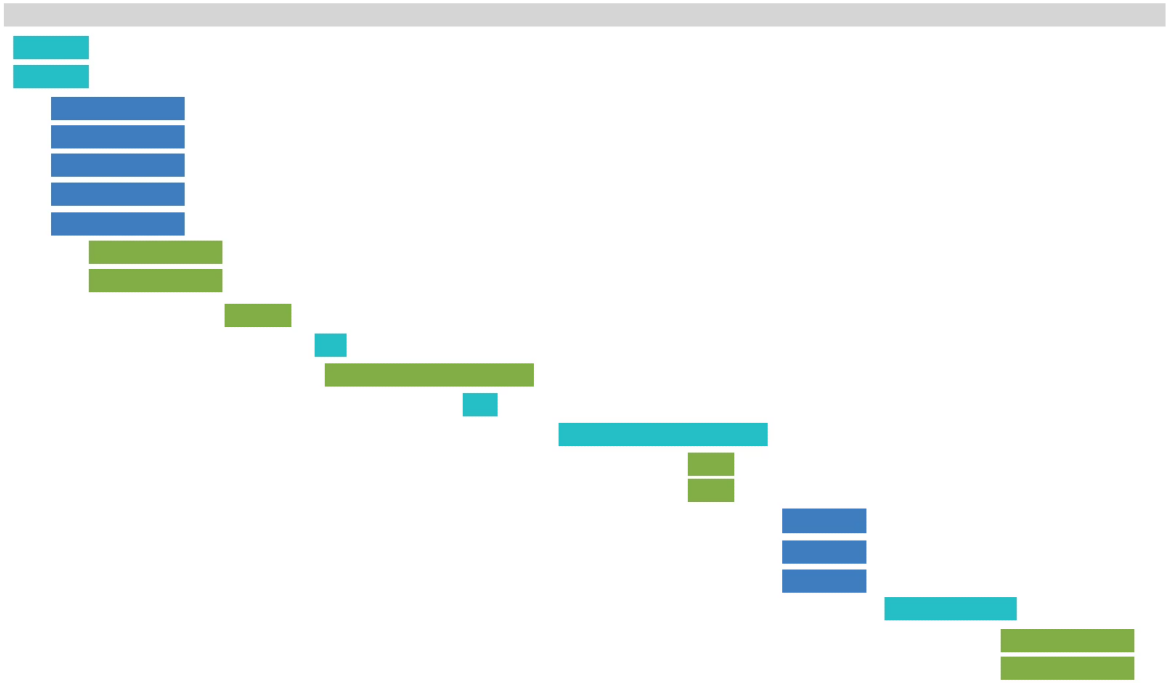
Position Sorting—Germline Pipeline Example

Mapping Aligning | **Position Sorting*** | Duplicate Marking* | Variant Calling



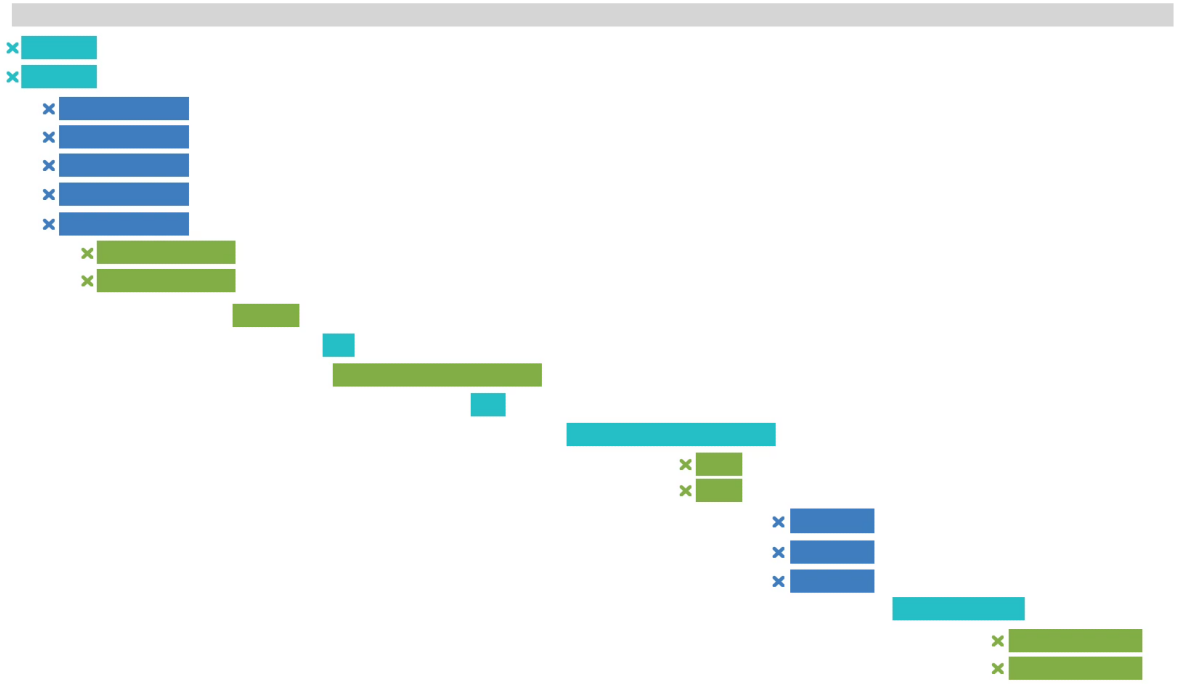
Duplicate Marking—Germline Pipeline Example

Mapping Aligning | Position Sorting* | **Duplicate Marking*** | Variant Calling



Variant Calling—Germline Pipeline Example

Mapping Aligning | Position Sorting* | Duplicate Marking* | **Variant Calling**



DRAGEN Automatic QC Metrics Reporting: Mapper

Removes the need to run downstream tools for QC

- Number of samples
- Reads Processed
- Total
- Biallelic
- Multiallelic
- SNPs
- INDELs
- MNPs
- SNP Transitions
- SNP Transversions
- Ti/Tv ratio

- Heterozygous
- Homozygous
- Het/Hom ratio
- In dbSNP
- Novel
- Total
- Biallelic
- Multiallelic
- SNPs
- INDELs
- MNPs

DRAGEN Automatic QC Metrics Reporting: Variant Caller

- Total input reads
- Number of duplicate reads (marked not removed)
- Number of unique reads
- Reads with mate sequenced
- Reads without mate sequenced
- QC-failed reads
- Mapped reads
- Number of unique & mapped reads (excl. dups)
- Unmapped reads
- Singleton reads (itself mapped; mate unmapped)
- Paired reads (itself & mate mapped)
- Properly paired reads
- Not properly paired reads (discordant)
- Reads with MAPQ [40:inf)
- Reads with MAPQ [30:40)
- Reads with MAPQ [20:30)
- Reads with MAPQ [10:20)
- Total reads in RG
- Supplementary (chimeric) alignments
- Average sequenced coverage over genome

- Total alignments
- Secondary alignments
- Supplementary (chimeric) alignments
- Estimated read length
- Bases in reference genome
- Bases in target bed [% of genome]
- Average sequenced coverage over genome
- Average alignment coverage over genome
- PCT of genome with coverage [40x:inf)
- PCT of genome with coverage [30x:40x)
- PCT of genome with coverage [20x:30x)
- PCT of genome with coverage [10x:20x)
- PCT of genome with coverage [5x:10x)
- PCT of genome with coverage [2x: 5x)
- PCT of genome with coverage [1x: 2x)
- PCT of genome with coverage [0x: 1x)
- DRAGEN mapping rate [mil. reads/second]
- Secondary alignments
- Estimated read length
- Insert length: mean

- Number of duplicate reads (marked)
- Number of unique reads
- Reads with mate sequenced
- Reads without mate sequenced
- QC-failed reads
- Mapped reads
- Number of unique & mapped reads (excl. dups)
- Unmapped reads
- Singleton reads (itself mapped; mate unmapped)
- Paired reads (itself & mate mapped)
- Properly paired reads
- Not properly paired reads (discordant)
- Reads with MAPQ [40:inf)
- Reads with MAPQ [30:40)
- Reads with MAPQ [20:30)
- Reads with MAPQ [10:20)
- Reads with MAPQ [0:10)
- Total alignments
- Secondary alignments
- Insert length: standard deviation

WGS Tools Help Genomic Epidemiology Characterize the Flow of Resistomes Between Animals and Humans

Bioinformatics tools available to detect Antimicrobial Resistance (AMR) determinants

- Public Health England's GeneFinder
- NCBI's AMRFinder
- KmerResistance
- SRST2
- ResFinder

WGS can predict AMR and the minimal inhibiting concentration (MIC) of an antimicrobial, applying machine or deep learning to genome sequence data[1].

NARMS
National Antimicrobial Resistance Monitoring System

FDA, the Centers for Disease Control and Prevention (CDC), and the U.S. Department of Agriculture (USDA) established the National Antimicrobial Resistance Monitoring System (NARMS) to systematically sequence pathogens from food-animals for AMR surveillance[2].

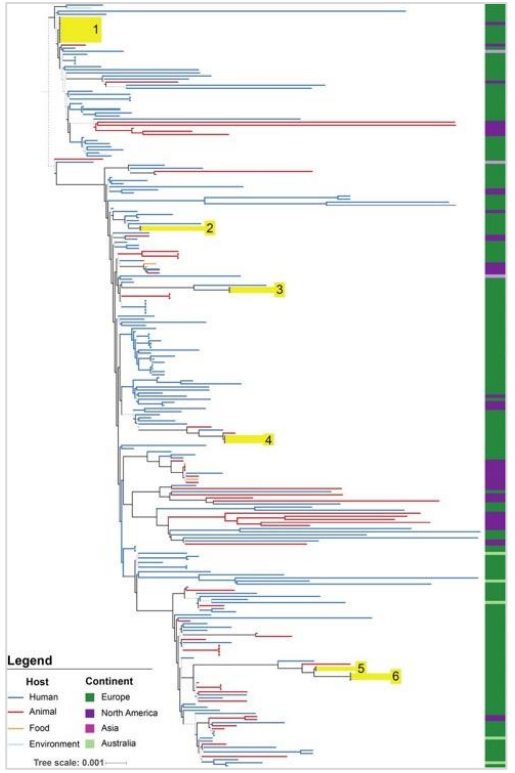
1. Scientific Report. (2018) 8:421–18972. doi: 10.1038/s41598-017-18972-w
2. Foodborne Pathog Dis. (2017) 14:545–57. doi: 10.1089/fpd.2017.2283

Culture-based Whole-Genome Sequencing (WGS) in Action

Zoonotic transfer of *Clostridium difficile* harboring antimicrobial resistance between farm animals and humans

"The phylogeny demonstrates clear mixing of European and North American strains, indicating multiple transmission events between continents, and mixing of human and animal strains, indicating multiple transmissions events between these hosts."

Sequenced on Illumina HiSeq™ Sequencer



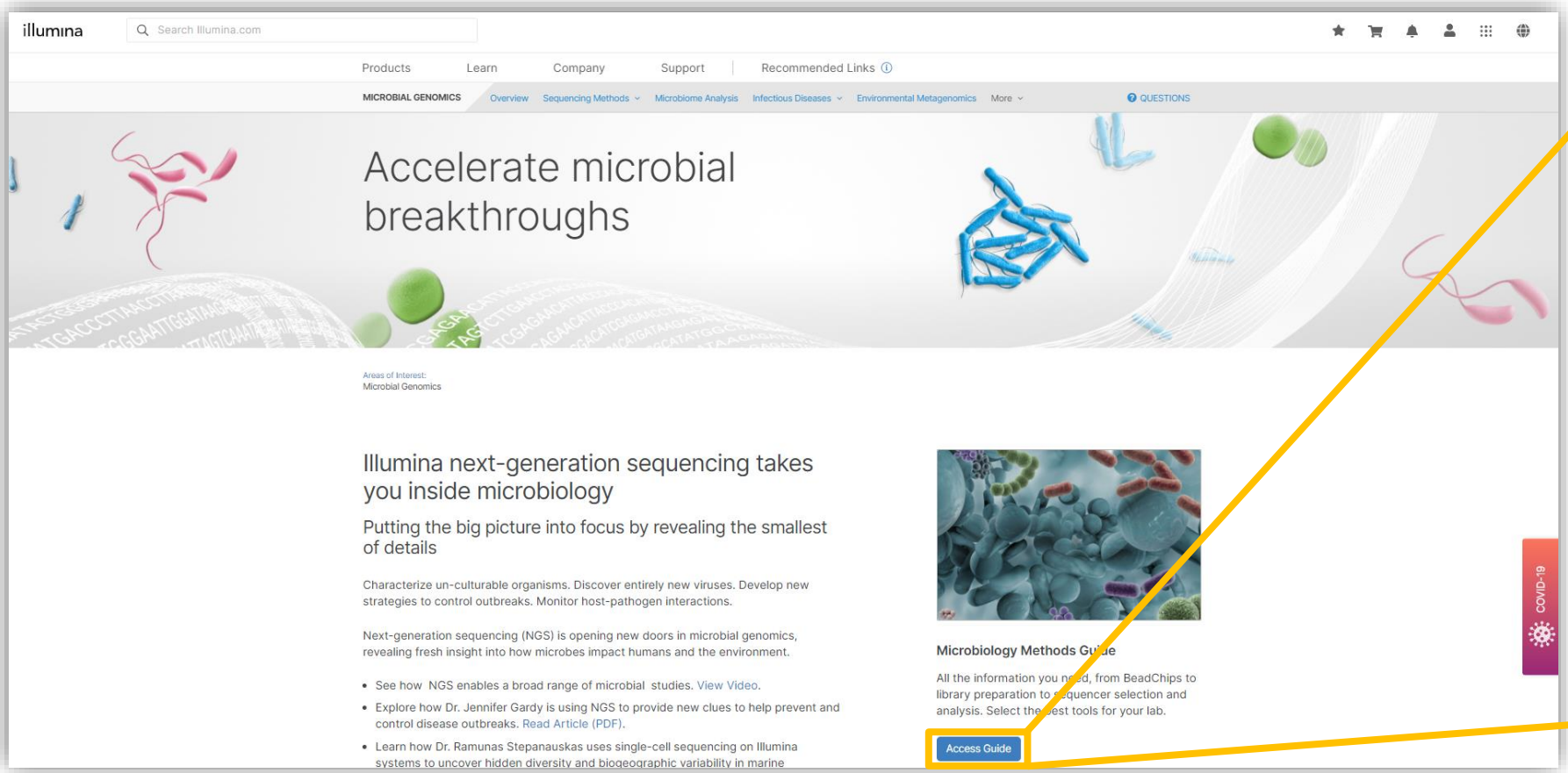
C. W. Knetsch et al. Journal of Clinical Microbiology Feb 2018, 56 (3) e01384-17; DOI: 10.1128/JCM.01384-17

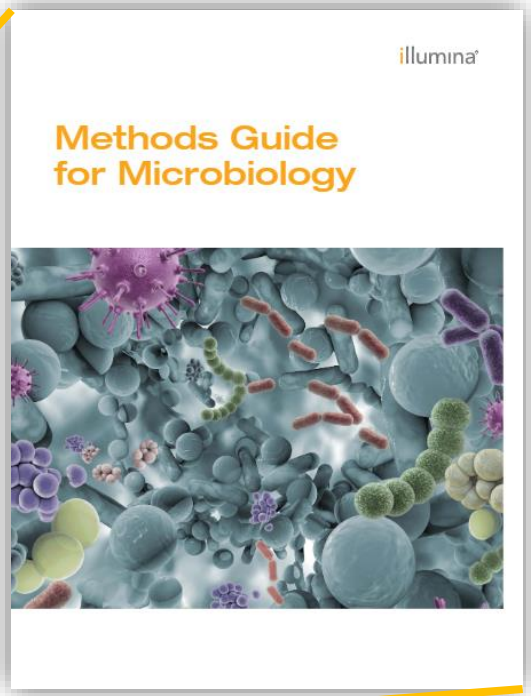
Microbial Genomics on the MiSeq

Advantage of Using Illumina Technology in Microbial Studies

- Sequencing the 5 Mb genome of *Escherichia coli* with Illumina technology can be done in one day at a fraction of the cost of traditional capillary electrophoresis methods
- Microbial genomes are small and data analysis is relatively simple
- NGS has the ability to measure the changes in the genome with prior knowledge
- Single-base resolution allows tracking of microbial adaptation over a short period of time

Microbial Genomics Resources





https://www.illumina.com/areas-of-interest/microbiology.html

Thank you

For any additional technical queries please contact Illumina Tech Support

techsupport@illumina.com or 1800.775.688