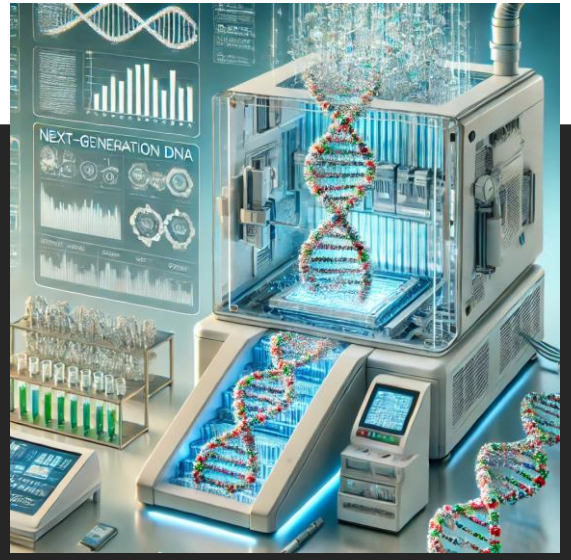
IABS NGS Training Workshop
December 3rd, 2024
Frankfurt, Germany

Introduction of NGS Bioinformatic Analysis



Pei-Ju Chin, M.S., Ph.D.

Division of Viral Product
Office of Vaccine Research and Review
Center for Biologics Evaluation and Research
U.S. Food and Drug Administration

Disclaimer

This presentation is an informal communication and represents my own best judgment. The material in this presentation and My comments an informal communication and represent my own best judgement. These comments do not bind or obligate FDA.

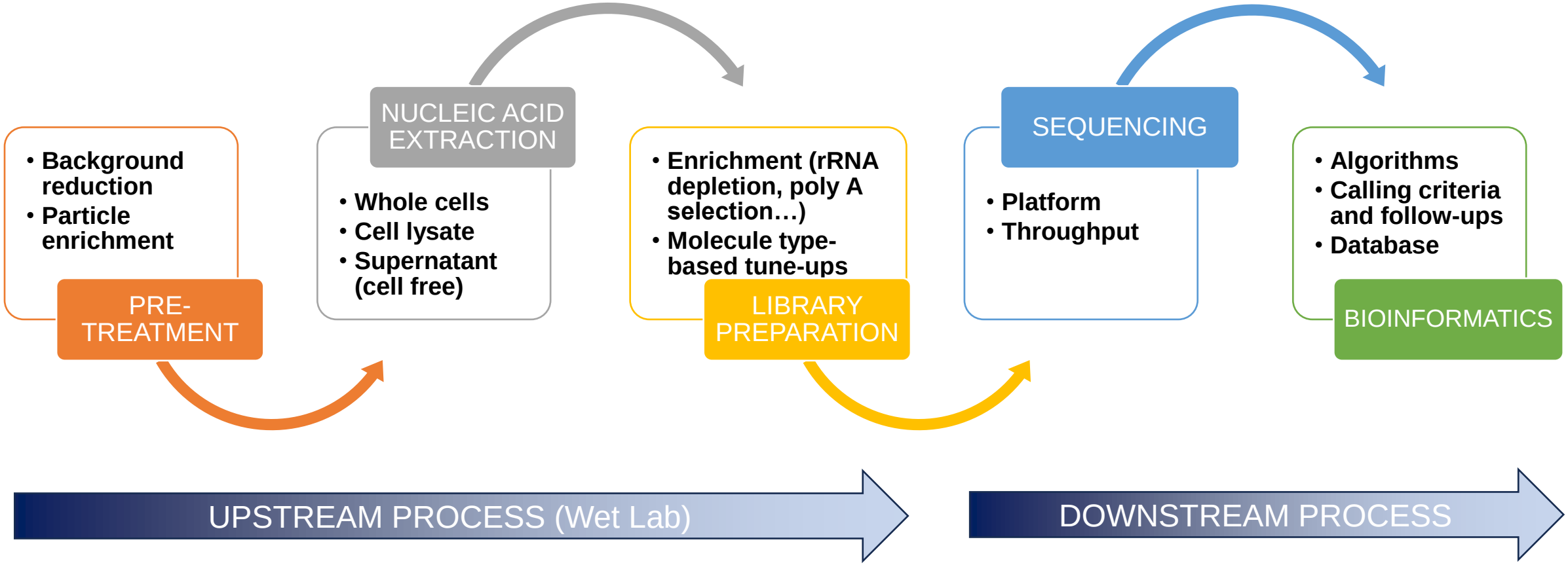
The concept provided in this presentation is simplified for the attendees without prior NGS analysis knowledge. The presenter encourages the attendees to further investigate the detail of NGS analysis algorithms, workflows and tools.

Outline

- **The concept of NGS Bioinformatics**
  - The goal of analysis and the methodology to achieve this goal
- **The overview of NGS Bioinformatic Pipeline**
  - Targeted and Non-targeted Analysis
  - Introduction of workflows
  - Bioinformatics resources
- **Demo**
  - **Unipro UGENE**

A General NGS Workflow

Wet Lab work is more important than in-silico work (garbage in, garbage out).
Bioinformatician can't fix all faults introduced in the wet Lab work.



Derived from Khan AS, and Study Group Participants. Proposed 1st International Virus Reference Standards for Adventitious Virus Detection in Biological Products by Next-Generation Sequencing (NGS)Technologies (CBER-5). Expert Committee on Biological Standardization. WHO/BS/2020.2394.

NGS Bioinformatics Is An Example of “Big Data” Science

You, as a data scientist, are performing data mining from billions of nucleotide sequences to extract meaningful information

Next Generation Sequencing – Platforms

Capacity
~ 6,5 Gb per day

Capacity
~ 3.000 Gb per day

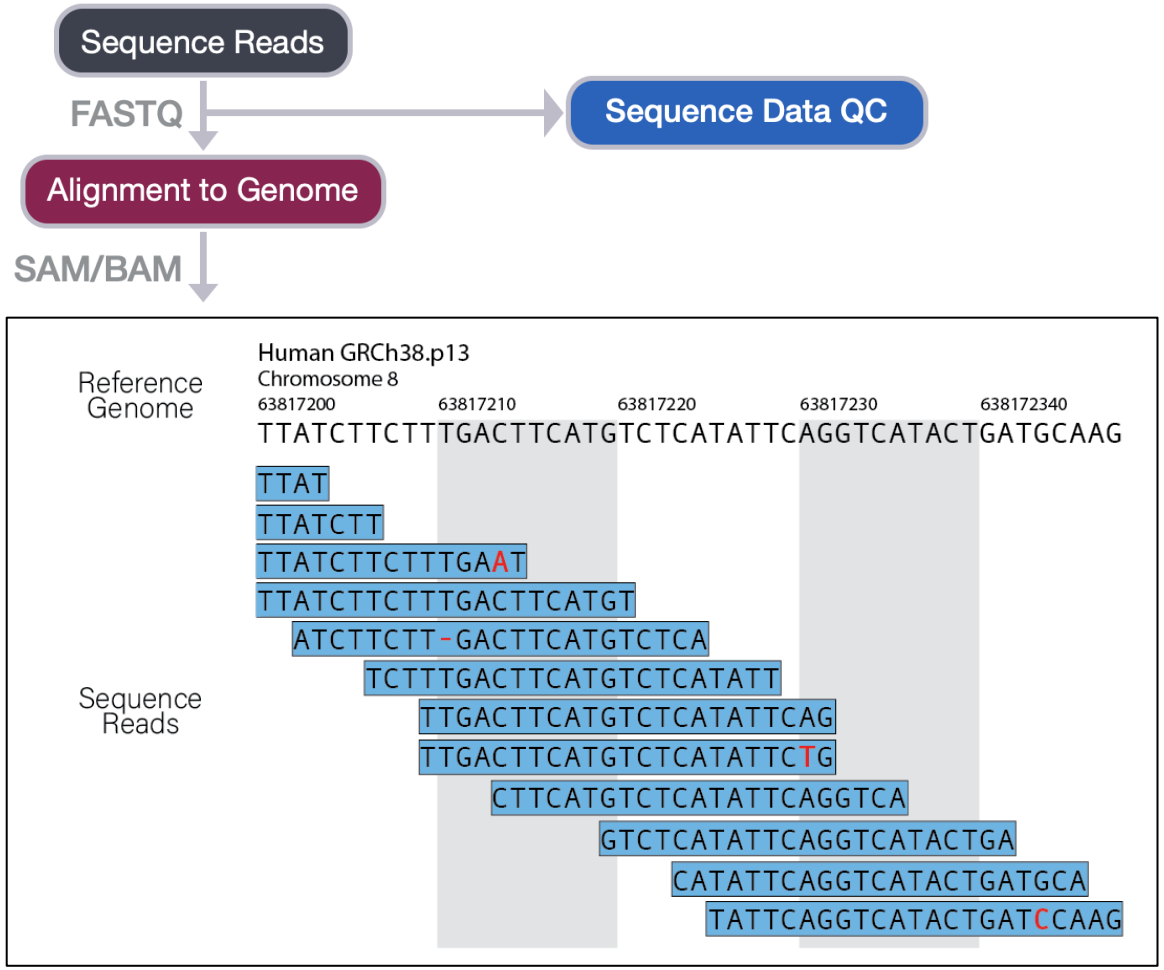
Capacity
~ 83 Gb per day

Sequencer	MiSeq	NovaSeq 6000	GridION
Manufacturer	Illumina	Illumina	Nanopore Technologies
Sequencing chemistry	Sequence-by-synthesis	Sequence-by-synthesis	Nanopore sequencing
Data output / run	13 - 15 Gb	4.800 – 6.000 Gb	2.8 - 50 Gb
Accuracy	99.9%	99.9%	99%
Time per run	~ 48 hours	16 - 44 hours	1 min - 72 hours
Read length	2 x 300 bp	2 x 150 bp	> 4 Mb

https://the-dna-universe.com/2021/11/25/how-to-do-ngs-50-faster-with-our-ngs-platforms/

NGS Bioinformatics: General Concept

Placing NGS reads to where they belong on the reference genomes

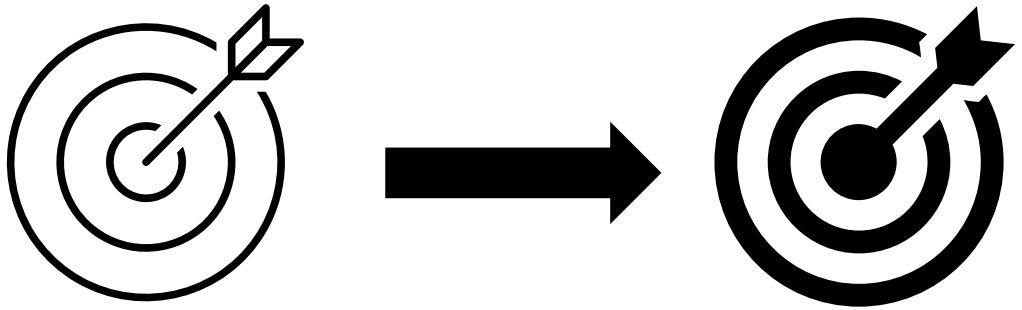


https://hbctraining.github.io/Intro-to-ChIPseq-flipped/lessons/04_alignment_using_bowtie2.html

NGS Analytical Approaches

Targeted Analysis

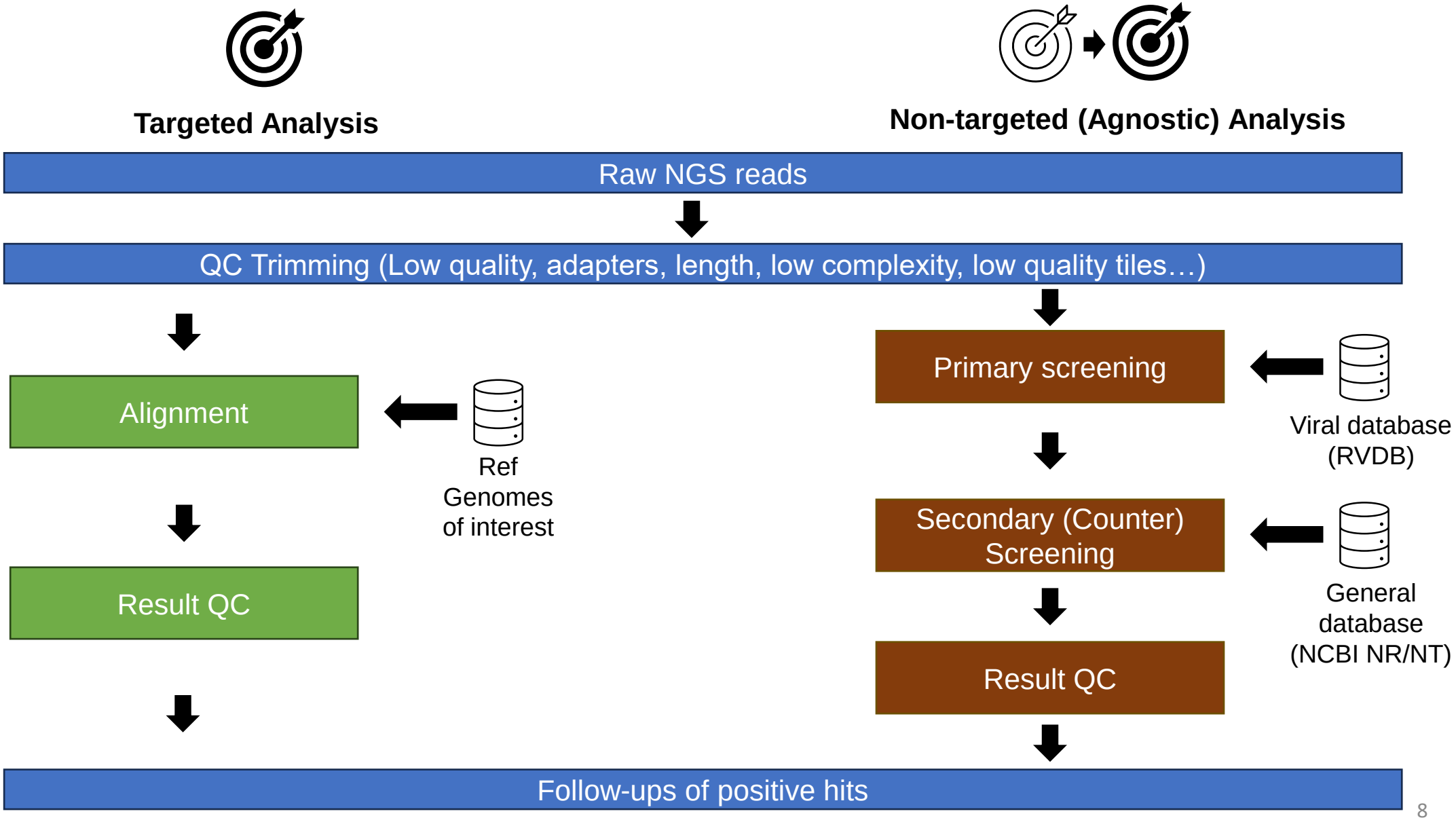
- ❑ You know what you are looking for (PCR as analogy)
- ❑ Suitable for identifying known/expected viruses to replace virus-specific PCR assays (ICH Q5A R2)
- ❑ Pro: Rapid detection, required less computational resources
- ❑ Con: Detection is limited to the reference genomes of known viruses



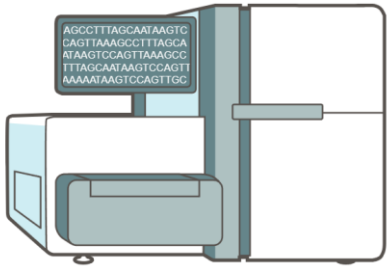
Non-targeted (Agnostic) Analysis

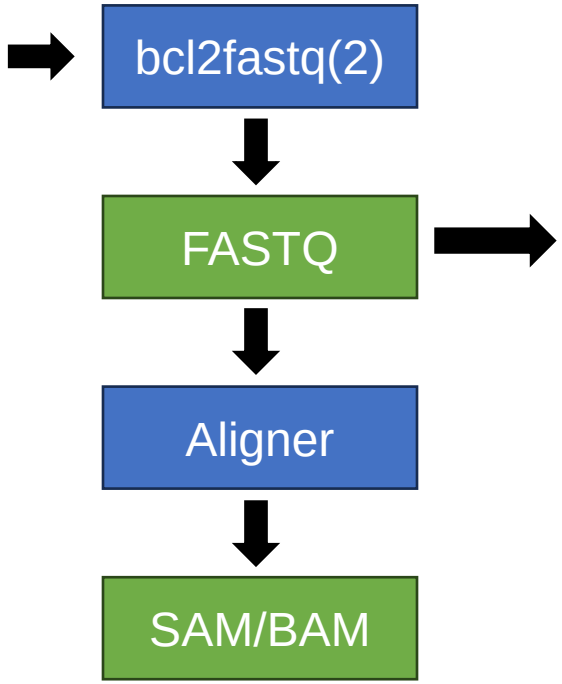
- ❑ You don’t know what you are looking for
- ❑ Suitable for detecting unknown adv agents to replace *in-vivo* assays (ICH Q5A R2)
- ❑ Pro: Broad detection (All viral sequences provided to the analysis pipeline)
- ❑ Con: Computationally-demanded

NGS Analytical Approaches: Generic Workflows



Essential Files in NGS Analysis Pipeline: FASTQ File



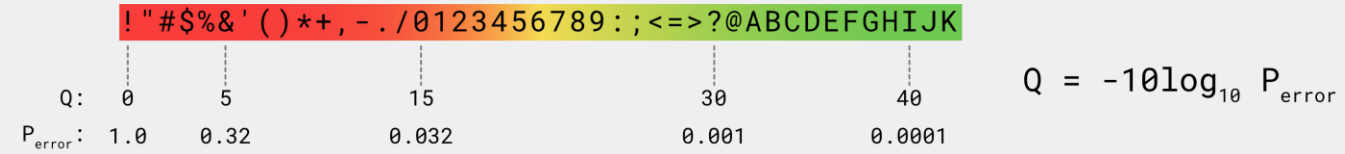


FASTQ File:

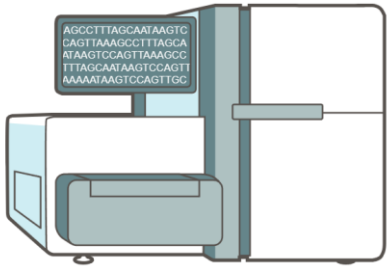
FASTQ file sample:

```
@SRR6407486.1 1 length=100
CCTCGTCTACAGCGACAACGTCCAGACCCGCGAACGGGTGATGCGGGCCCTGGGCAAACGGTTGCACCCGGATCTGCCCGATTTGACCTACGTCGAAGTG
+SRR6407486.1 1 length=100
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF<FFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFF7FFFF<FF
```

Quality scores as ASCII characters:



Essential Files in NGS Analysis Pipeline: SAM/BAM File



bcl2fastq(2)

FASTQ

Aligner

SAM/BAM

Sequence Alignment Map (SAM)/Binary Alignment Map (BAM)

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Header section

Alignment section

Optional fields in the format of TAG:TYPE:VALUE

QUAL: read quality; * meaning such information is not available

SEQ: read sequence

TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.

PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.

RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.

CIGAR: summary of alignment, e.g. insertion, deletion

MAPQ: mapping quality

POS: 1-based position

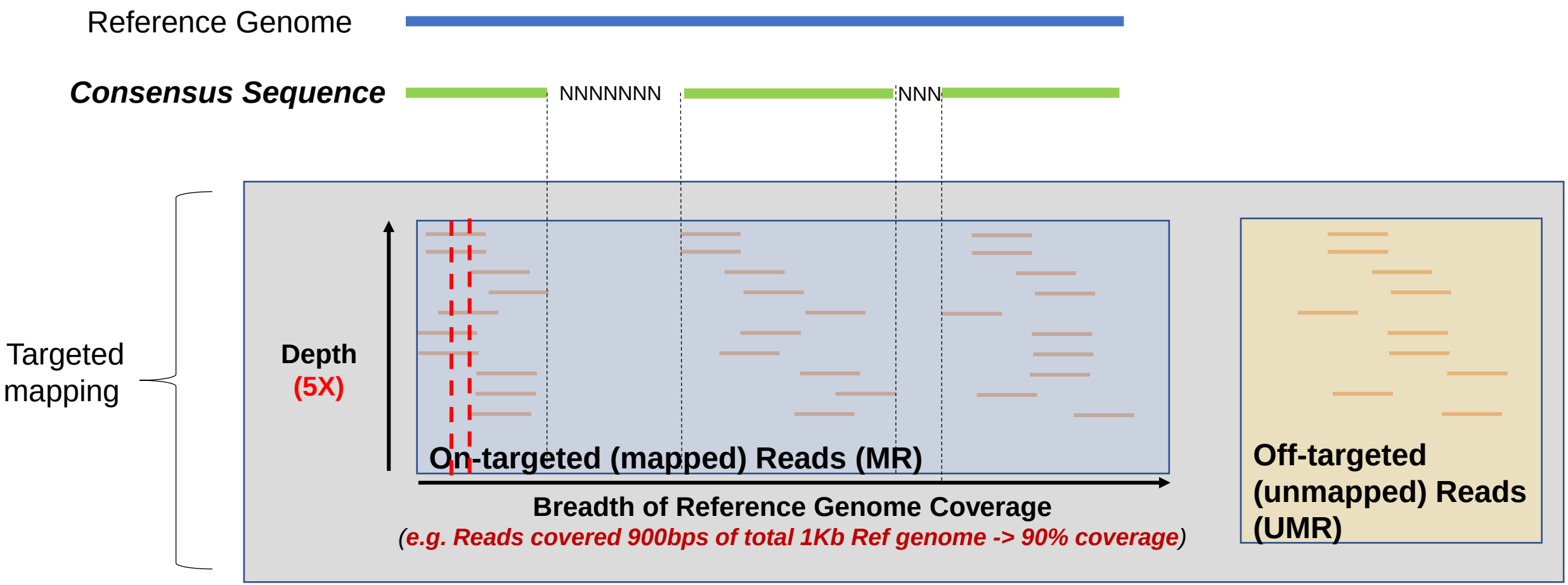
RNAME: reference sequence name, e.g. chromosome/transcript id

FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.

QNAME: query template name, aka. read ID

Terminology of NGS Mapping Matrices

Consensus sequence, sequencing depth, genome coverage, mapped and unmapped reads



NGS Bioinformatics: Resources

- ***Online analysis platforms***
  - Galaxy https://usegalaxy.org/
  - Chan Zuckerberg ID https://czid.org/
- ***Tools***
  - Qiagen CLC Genomics Workbench https://digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/analysis-and-visualization/qiagen-clc-genomics-workbench/
  - Geneious https://www.geneious.com/
  - DNANexus https://www.dnanexus.com/use-cases/genomic-analysis
  - GATK https://gatk.broadinstitute.org/hc/en-us
  - UGENE https://ugene.net/
- ***Databases***
  - NCBI https://www.ncbi.nlm.nih.gov/
  - Reference Virus Database (RVDB) https://rvdb.dbi.udel.edu/home

Demo

- Targeted analysis by UniPro UGENE https://ugene.net/
- NGS reads
  - Subsampled FASTQ Paired-end FASTQ files from one of the spiking studies conducted in Khan Lab
- Reference genomes
  - EBV, PCV1, MVM, FeLV, hCoV OC43, RSV, REO, Adenovirus Type 5, SMRV, PERV
- Scenario
  - To detect if any 11 viruses are presented in the sample.
  - Follow-ups of candidate hits by NCBI BLAST

Supplemental Slides

NGS Bioinformatics: Terminology