4th Conference on Next Generation Sequencing for Adventitious Virus Detection in Biologics for Humans and Animals: Validation and Implementation of NGS

December 3-5, 2024

Frankfurt, Germany

Considerations for the Follow-Up of Putative Adventitious Virus Signals Detected by High-throughput Sequencing (HTS)

Christophe Lambert

Disclaimer

- Christophe Lambert is an employee of the GSK group of companies.
- The content and views expressed in this presentation are derived from the individual experts participating to the AVDTWG Subgroup DE and should not be construed to represent the views or policy of the organization, regulatory authority, or governmental agency they represent.
- All criteria, tools or ways of working described must be considered as “considerations” only and should not be taken as recommendations, mandatory use or regulatory views.

Adventitious agent testing

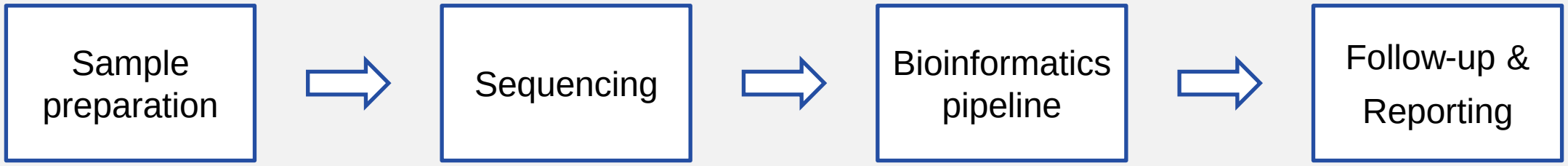
Adventitious agent tests are routinely used to assess safety and purity of cell banks and biologics, and include:

- Electron microscopy
- Assays for retroviral reverse transcriptase
- *In vivo* assays – specific animal species inoculated with samples; detection based on:
  - mortality
  - testing of tissue for the presence of hemagglutinins
- *In vitro* assays – samples applied to multiple cell lines; detection based on:
  - cytopathic effect
  - hemagglutination
  - hemadsorption
- *In vitro* assays – virus-specific (q)(RT-)PCR

Adventitious virus detection using HTS

- HTS demonstrated capabilities for broad virus detection
  - discovery of known and novel viruses in a variety of samples, including clinical, environmental, and biological

- The typical workflow for virus detection using HTS involves a combination of laboratory and bioinformatics steps:



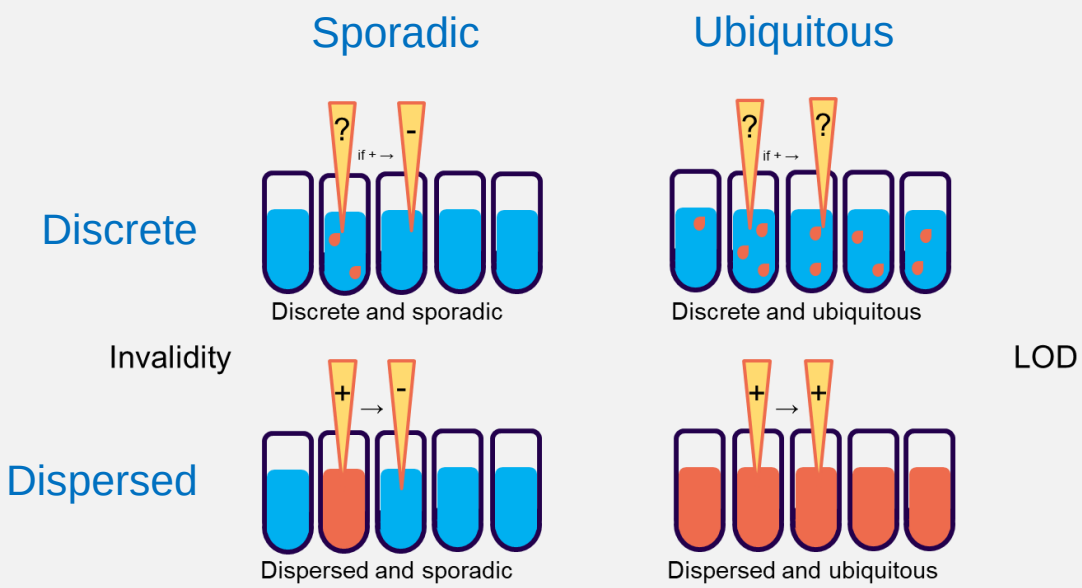
- There are key challenges in each step of the workflow

Potential sources of false-positive signals

- Human shedding during testing or within test reagents
  - human commensals and their bacteriophages; human viruses
- Laboratory cross-contamination and environmental contamination (during testing or within test reagents)
  - a variety of organisms and their viruses, equipment contamination, laboratory surfaces or air
  - viruses and organisms used in other experiments, index switching and barcode carryover
- Residuals within analytical reagents resulting from their production
  - host cell endogenous viruses; *E. coli* and their bacteriophages; viral cloning and expression vectors
- Viral sequences integrated into the host genome
  - constructs used in engineering cell lines, plus endogenous viruses
- False-positive bioinformatic identification
  - misidentification due to sequencing errors, natural variation or sequencing artifacts
  - tentative identifications, such as with weak matches
  - incomplete public databases, or misannotated database entries

Source of sample variability

- Full versus segmented genomes
- Natural variations in viral genomes / population
- Sampling short reads / long reads / assemblies
- Inequal distribution of reads (stacked reads)
- Evolutionary divergence
- Discrete versus dispersed
- Sporadic versus ubiquitous

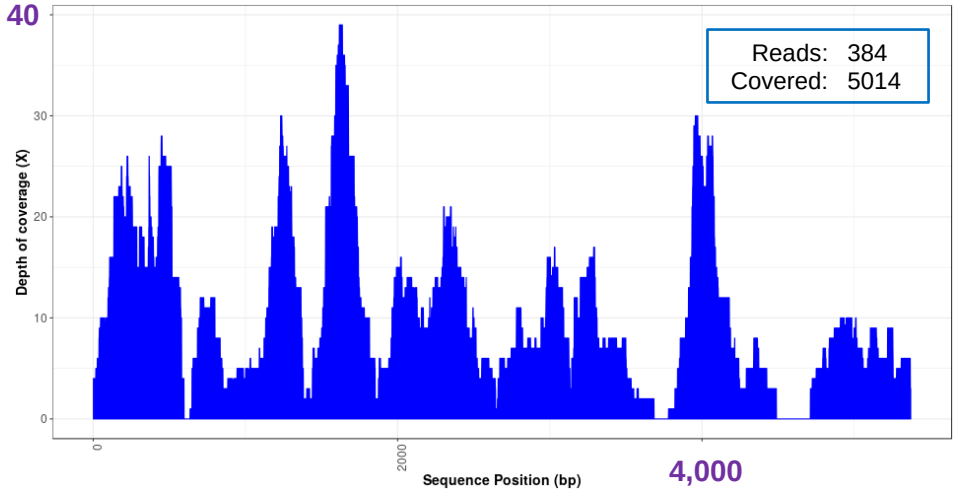


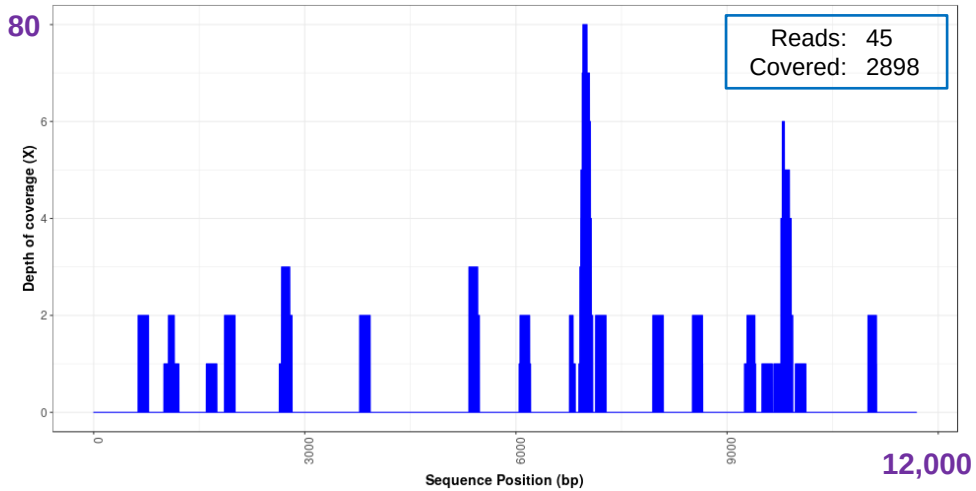
Inequal distribution of reads (stacked reads)

Typical

True positives

Reads spread over the whole genome



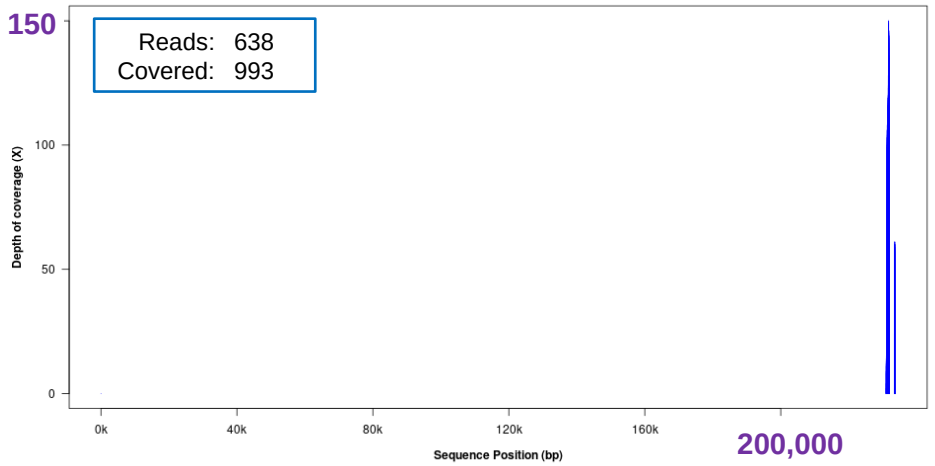


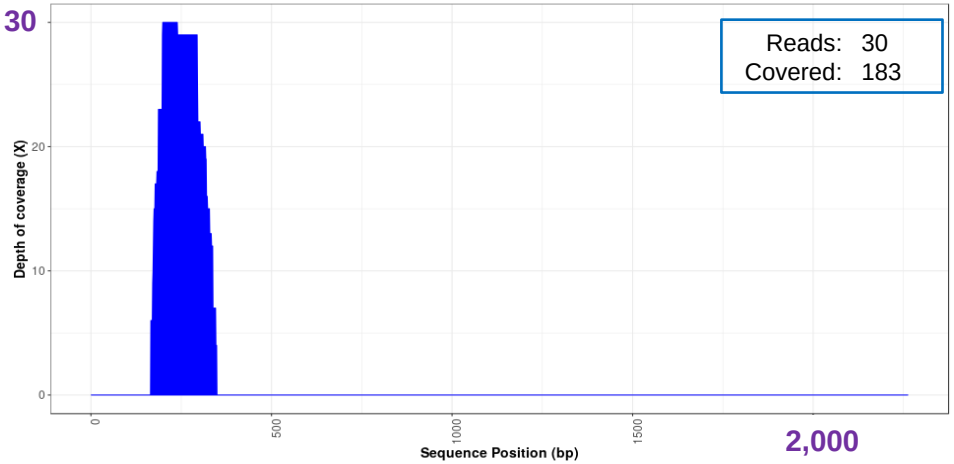
Typical

False positives

Stacked reads

Read length: 151 nt



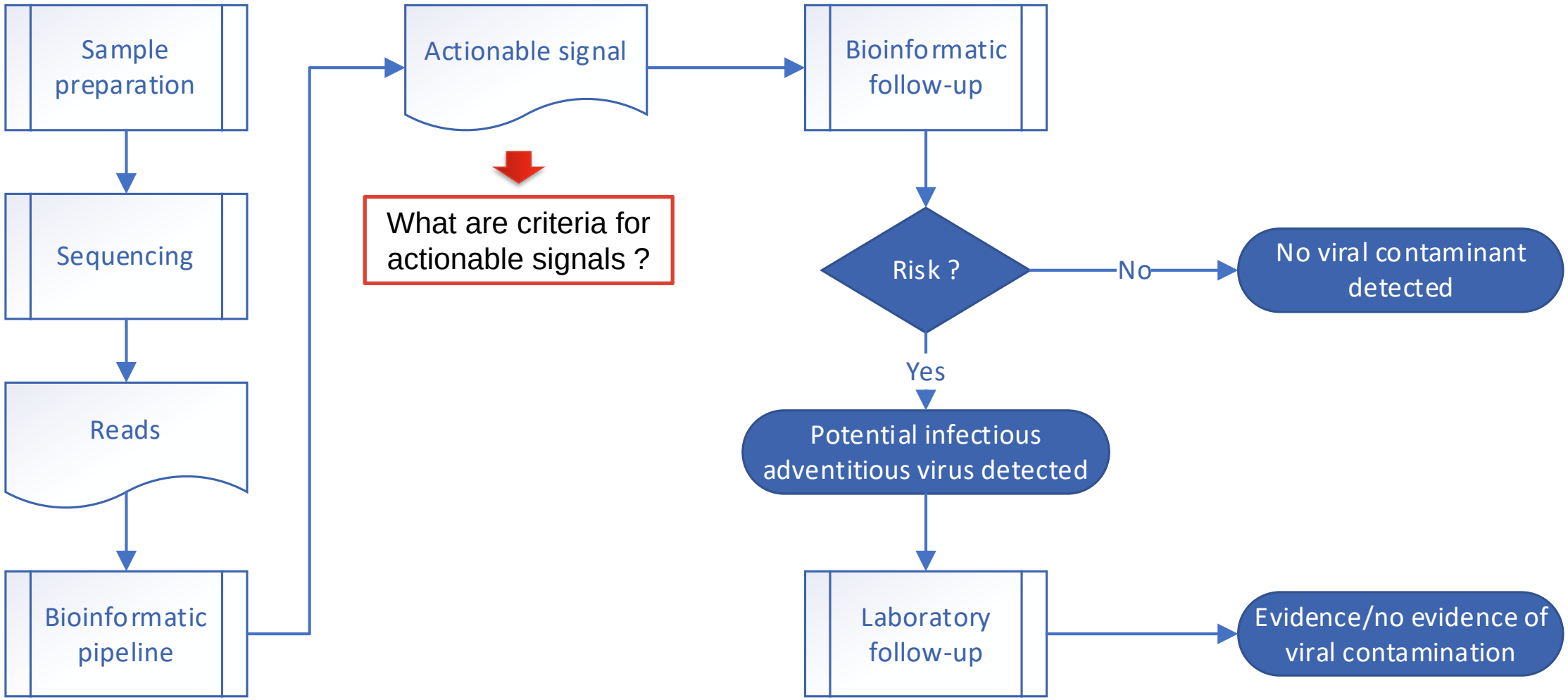


Reference Database Selection

- Databases used at different steps of the bioinformatic pipeline
  - screening for potential viral reads
  - counter-screening of potential viral reads

- Considerations for database selection
  - completeness of the database
  - quality of sequences (presence of Ns)
  - quality of the annotation (chimeric and contaminated sequences, unidentified, environmental samples, uncultured virus…)
  - representation of taxonomies
  - up-to-date and maintained database
  - addressing the knowledge gaps in taxonomy from metagenomics sequences

Follow-up of HTS-based adventitious virus detection



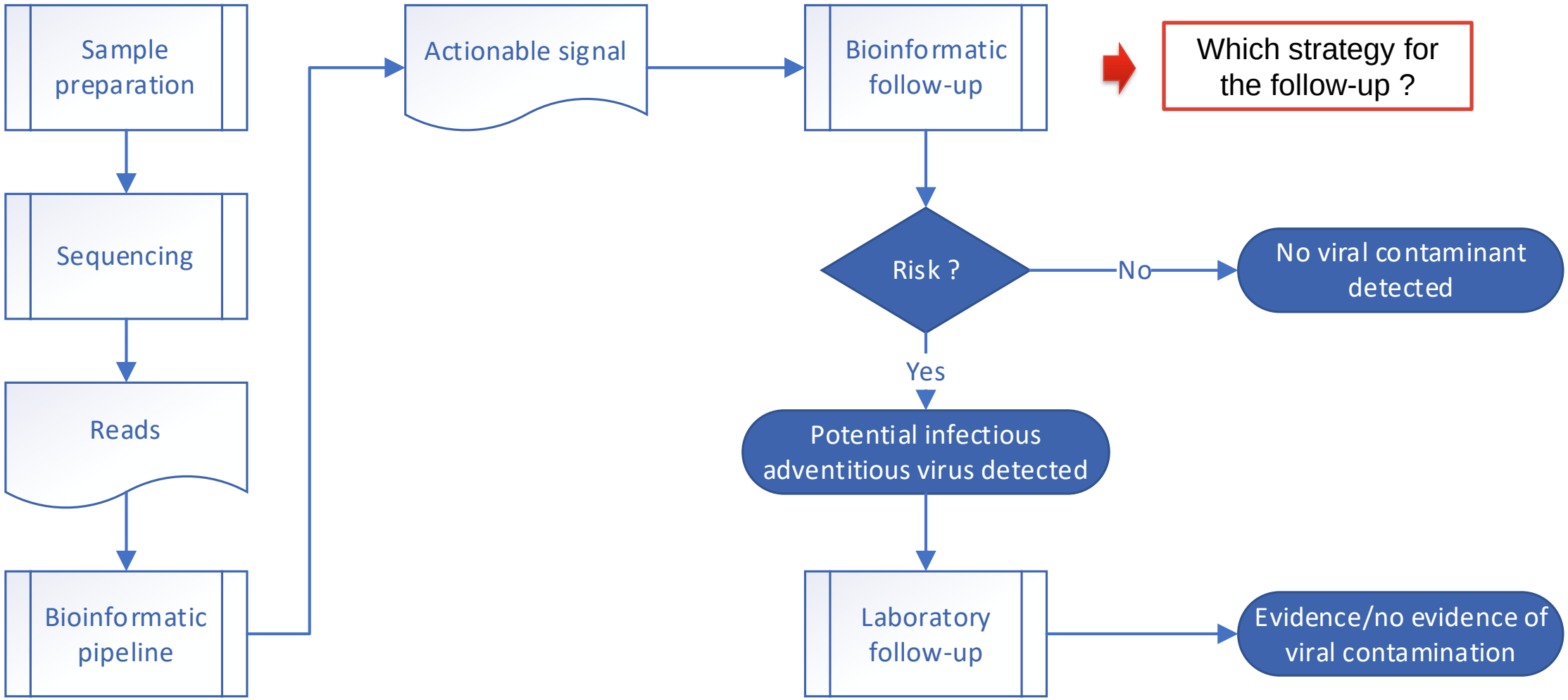
Criteria for consideration of an actionable signal

Criterion	Rationale to choose a value
Number of reads (function of total reads and size of genome)	Arbitrary† threshold; independent molecules for confirmation
Strength of match (function of sequence length and % identity)	Arbitrary† threshold; identity difficult to ascertain if the match is weak
Read alignment (function of % sequence identity)	Arbitrary† threshold; guard against chimeric constructs
Genomic coverage (function of read #, % identity)*	Using statistical and evolutionary models, including machine learning†
Taxonomy (function of % sequence identity)	Risk assessment for impact on cell lines, patients, and product quality
Quality assurance/control (function of 6σ – 5M1P)	Using experience, positive/negative controls, prior knowledge or machine learning†

* and in some cases, patterns of gene expression; different for total nucleic acid versus transcriptomic analyses; limited in e.g., latent expression
† refined by experience, including limit of detection studies and previous follow-up investigations

But criteria are interdependent, multidimensional... and signal dependent…

Follow-up of HTS-based adventitious virus detection



Considerations for bioinformatic/SME follow-up

Investigation	Information
Genome completeness	• Are identified sequence(s) complete viral genomes ? • If not, are there other identified sequence(s) corresponding to other viral regions ?
Non-coding region coverage	• Assess whether non-coding regions are covered by reads.
Similarity investigation / quality	• Confirm the identified viral region belongs to the annotated virus species. • Is the identified viral region similar to host sequences or other species (typically bacteria) ?
Characterize host / viral signal from other sources	• Are identified viral species/regions also found in available internal data (prior experiment) or public data for the same or similar matrix ?
Index Quality Assessment	• Are indexes of reads assigned to the identified viral sequence enriched in mismatches ? • Were Unique Dual Indexes (UDI) used to reduce the risk of index hopping ?
Collect information about the identified virus	• Literature search. What is this virus, its geographical distribution, its replication cycle, its host(s)… ?

As much information as possible is needed to decide on potential infectious adventitious virus detection.

Bioinformatic/SME follow-up (1)

Examples of actions

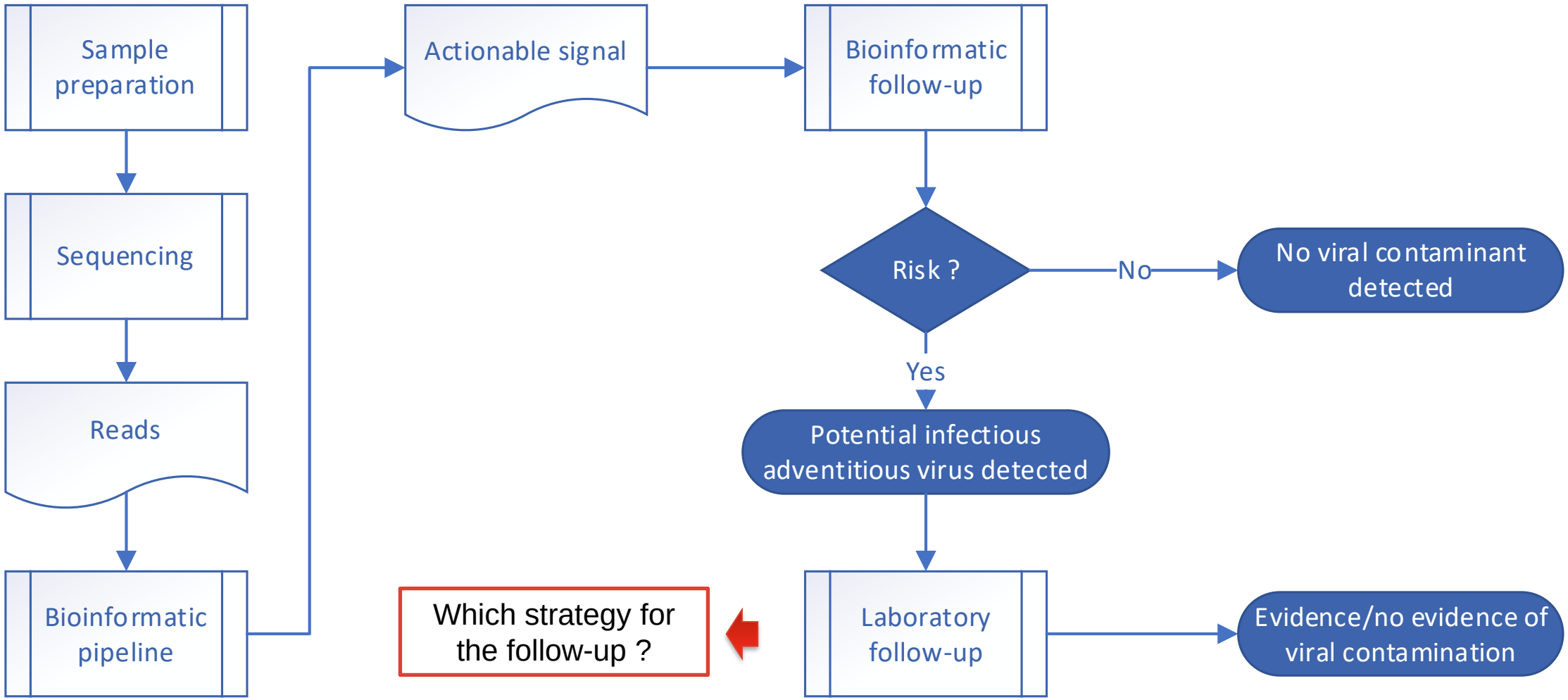
- Genome completeness
  - ➢Check whether the detected sequence corresponds to a complete genome in the NCBI database (https://www.ncbi.nlm.nih.gov/nuccore/). It could be deduced from its description and its length.
  - ➢If the sequence represents a fragment (e.g., one gene), verify in the results if other covered accessions correspond to the remainder of the virus.
- Coding/non-coding region coverage
  - ➢Coding and non-coding regions are sometimes annotated in the NCBI database.
  - ➢For DNA viruses, observing RNA coverage in coding regions could demonstrate that the virus is active.
  - ➢For RNA viruses, small RNA coverage only in coding regions could result from similar sequence in a eukaryotic genome in the sample, potentially from the host. Differential abundance of reads between coding and non-coding regions could demonstrate that the virus is active.
- Quality assessment
  - ➢Reads aligned to viral sequences can be reviewed based on specific metrics (percentage of identity, alignment breadth of coverage, sequence complexity, number of Ns, number of exact matches, alignment with the same species) (see Criteria for consideration of an actionable signal).

Bioinformatic/SME follow-up (2)

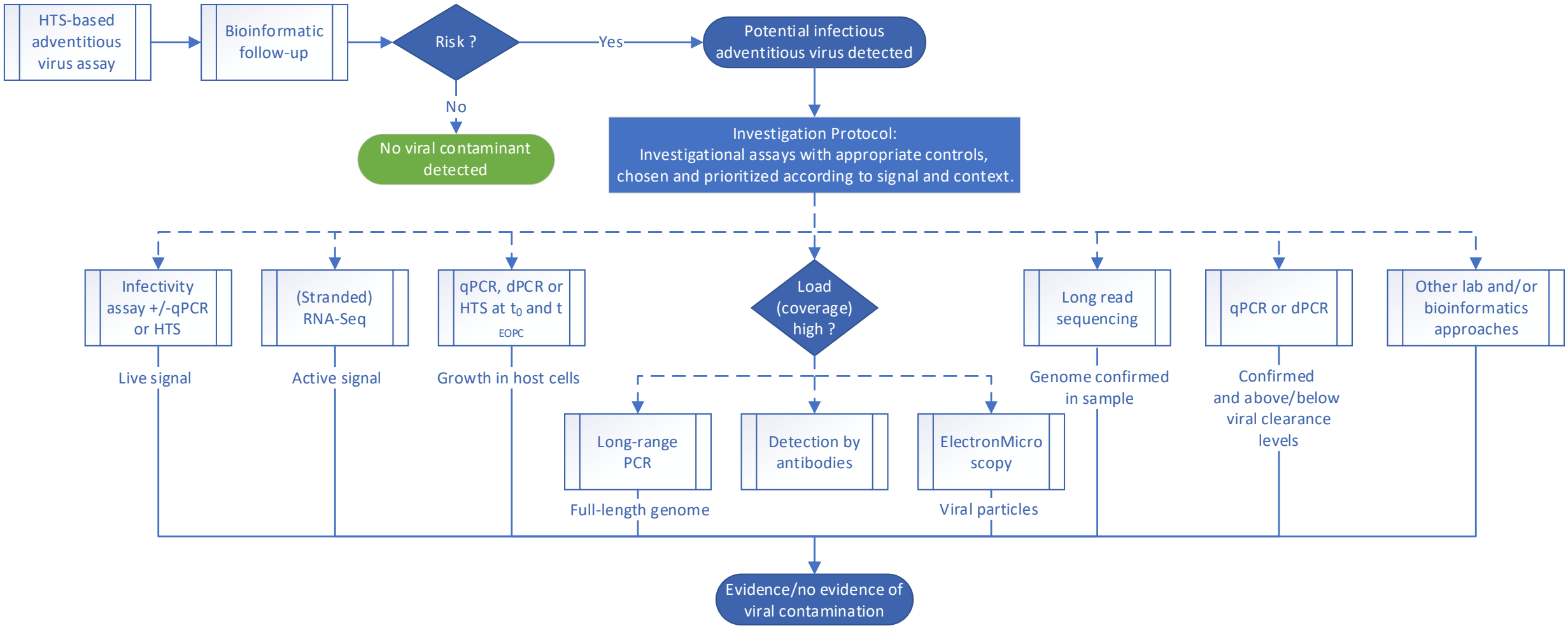
Examples of actions

- Similarity investigation
  - ➢The covered region of the detected viral sequence could be aligned to RefSeq/NR/GenBank using BLAST to assess whether the covered region also aligns with the host or other organisms.
  - ➢Reads assigned to viral sequence/species could be aligned to RefSeq/NR/GenBank using BWA/Bowtie/Minimap2 to assess whether they also align with the host or other organisms.
  - ➢Similarity investigation can support the actual presence of viruses or identify potential false positives.
- Characterize host / viral signal from other sources
  - ➢If sequencing data are available internally or publicly (typically SRA) for samples similar to the tested sample, analyzing those datasets can help determine whether the unexpected viruses detected are unique to the tested sample.
- Index Quality Assessment
  - ➢**Mismatches in indexes**: Investigation to determine if mismatches in index sequences could have resulted in reads being incorrectly attributed to a sample.
  - ➢**Distribution of index pairs**: Investigate if each multiplexed sample had a unique pair of indexes, referred to as Unique Dual Indexes (UDI), and that no index from a pair is shared with any other sample in the experiment to minimize index hopping.

Follow-up of HTS-based adventitious virus detection



Laboratory follow-up strategy



Acknowledgements

All AVDTWG Subgroup DE members

Contributors to the paper in preparation

Andrew S. McKay (Genentech)
Arifa Khan (US FDA/CBER)
Aurash Mohaimani (Biogen)
Jakob M. Goldmann (Merck)
Robert Charlebois (Sanofi)
Vanessa V. Sarathy (Merck & Co. Inc.)
Keishin Sugawara (Kumamoto University)
Nasrin Salehi (Pfizer Inc.)
Pei-Ju Chin (US FDA/CBER)
Jerry H. Lo (Genentech)
Simone Olgiati (Merck-Serono)

GSK

Olivier Vandeputte
Noémie Deneyer
Anne-Sophie Colinet

On behalf of GSK

Sophie Ayama
Léo d'Agata

Backup slides

Advanced Virus Detection Technologies WG

Mission: To facilitate the use of advanced technologies (such as HTS/NGS) for the detection of adventitious viruses in biologics by providing an informal, scientific forum for knowledge exchange, scientific discussions and collaborations among scientists across different organizations.

Organization:

> 200 participants

> 60 organizations

- *Regulatory agencies*
- *Government agencies*
- *Industries*
- *Service providers*
- *Technology developers*
- *Academia*

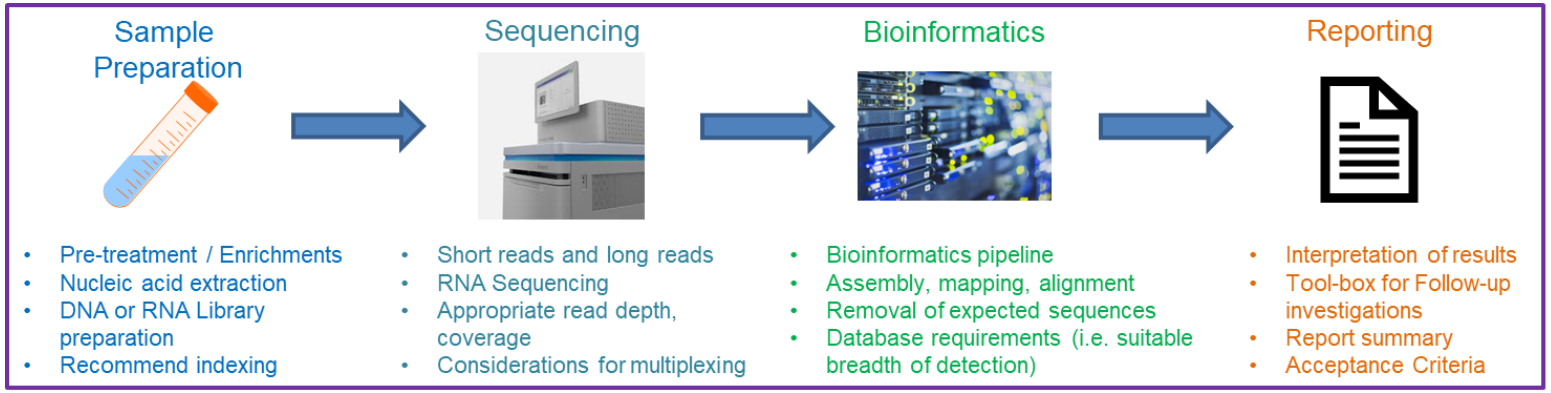
Meeting/discussions virtually once every 2 months

Co-chairs:

Arifa S. Khan (FDA/CBER, U.S.A.; October, 2012)
Siemon Ng (Notch Therapeutics, Canada; June, 2022)
Ken Kono (National Institute of Health, Japan; October, 2023)
Noémie Deneyer (GSK, Belgium; November, 2023)

AVDTWG Subgroups

Multiple key challenges requiring additional scientific data



2012

Subgroup A

Sample selection/ preparation/processing

Subgroup B

Virus standards and reference materials

Subgroup C

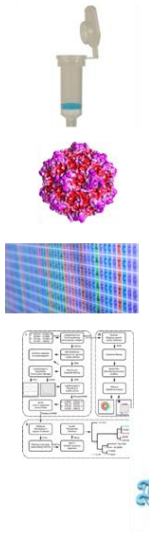
Complete and correctly annotated, virus reference database

Subgroup D

Bioinformatics pipelines analysis

Subgroup E

Follow-up strategies to confirm the identity of a "hit"



2024

Subgroup AB

Subgroup C

Subgroup DE

Co-leaders:

- *Christophe Lambert (GSK)*
- *Robert Charlebois (SANOFI)*

Organization:

~ 100 participants

~ 40 organizations

- *Regulatory & Government agencies*
- *Industries*
- *Academia*

AVDTWG Subgroup DE

Leader(s)

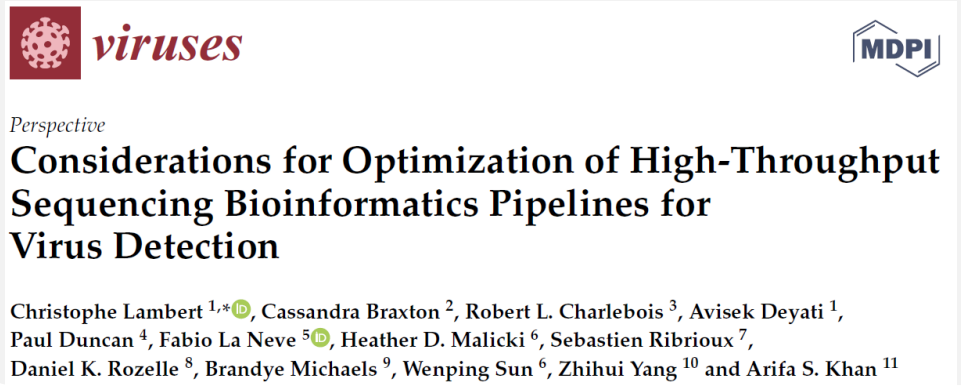
- Robert CHARLEBOIS
- Christophe LAMBERT

Advanced Virus Detection Technologies Working Group (AVDTWG)

- To facilitate the use of advanced technologies for the detection of adventitious viruses in biologics by providing an informal, scientific forum for knowledge exchange, scientific discussions and collaborations among scientists across different organizations.

AVDTWG Subgroup DE

- Optimization of Bioinformatics pipelines for adventitious virus detection and Strategy for follow-up investigations of identified hits.



- **Factors Influencing Sensitivity of Virus Detection**
  - Upstream Preparation of the Biological Sample
  - Sequencing
  - Bioinformatics Pipeline and Databases
- **Sequencing Platform and Output Files**
- **Data Analysis Pipeline Design**
  - Sequence Read Pre-Processing, Assembly, and Alignment
  - Database Selection
  - Reference Subtraction and Counter-Screen
  - Processing of Viral Hits
  - Unmapped Sequences
- **Data Captured in Raw Output**
- **Final Reporting Format**

Bioinformatics pipeline for virus detection using HTS

1. Quality control of raw reads

2. Assembly / no assembly

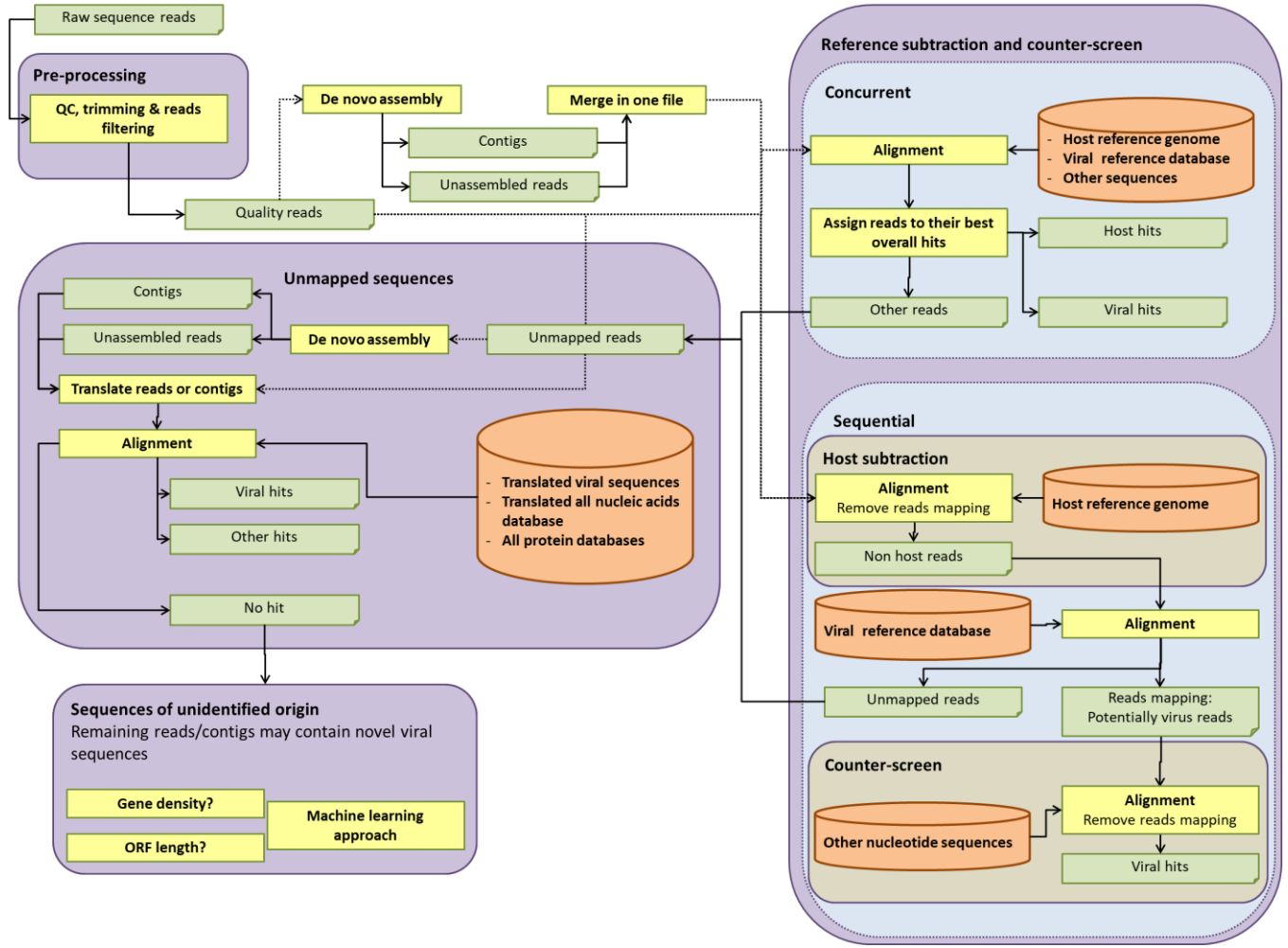
3. Host removal

4. Virus screening

5. Counter-screening

6. Treatment of unmapped sequences

7. Post-analysis of actionable signals



Lambert et al. Considerations for Optimization of High-Throughput Sequencing Bioinformatics Pipelines for Virus Detection. Viruses. 2018 Sep 27;10(10):528.