sanofi

Evolution of a bioinformatics pipeline for adventitious agent detection

~ how to find the proverbial needle in a haystack ~

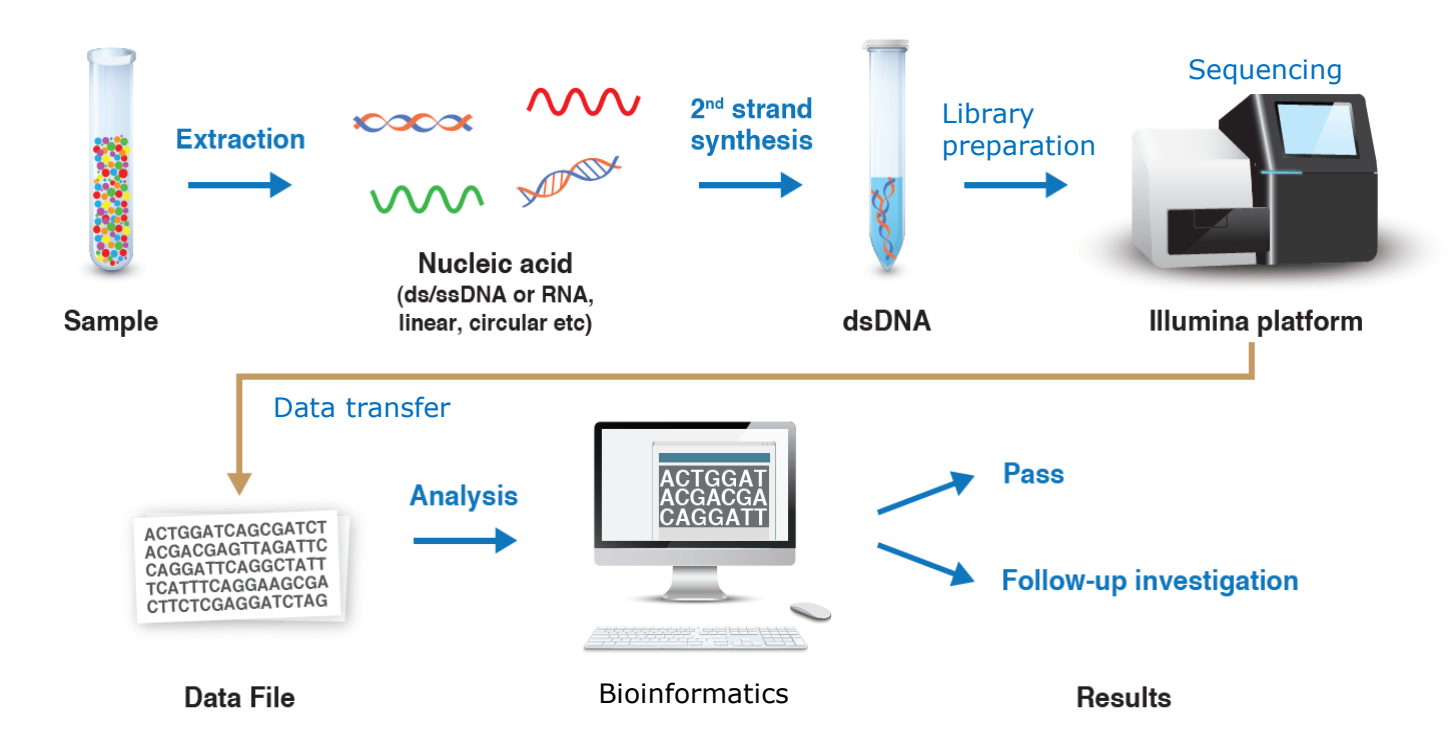
presented by Robert L. Charlebois

Head of the Molecular Biology Centre,
Analytical Sciences, Sanofi Vaccines, Toronto

IABS
05 Dec 2024

High-throughput sequencing for adventitious agent detection

Modular approach to facilitate upgrades and validation



The design space for the bioinformatic detection of signals

viruses

MDPI

Perspective

Considerations for Optimization of High-Throughput Sequencing Bioinformatics Pipelines for Virus Detection

Christophe Lambert [1,*], Cassandra Braxton [2], Robert L. Charlebois [3], Avisek Deyati [1], Paul Duncan [4], Fabio La Neve [5], Heather D. Malicki [6], Sebastien Ribrioux [7], Daniel K. Rozelle [8], Brandye Michaels [9], Wenping Sun [6], Zhihui Yang [10] and Arifa S. Khan [11]

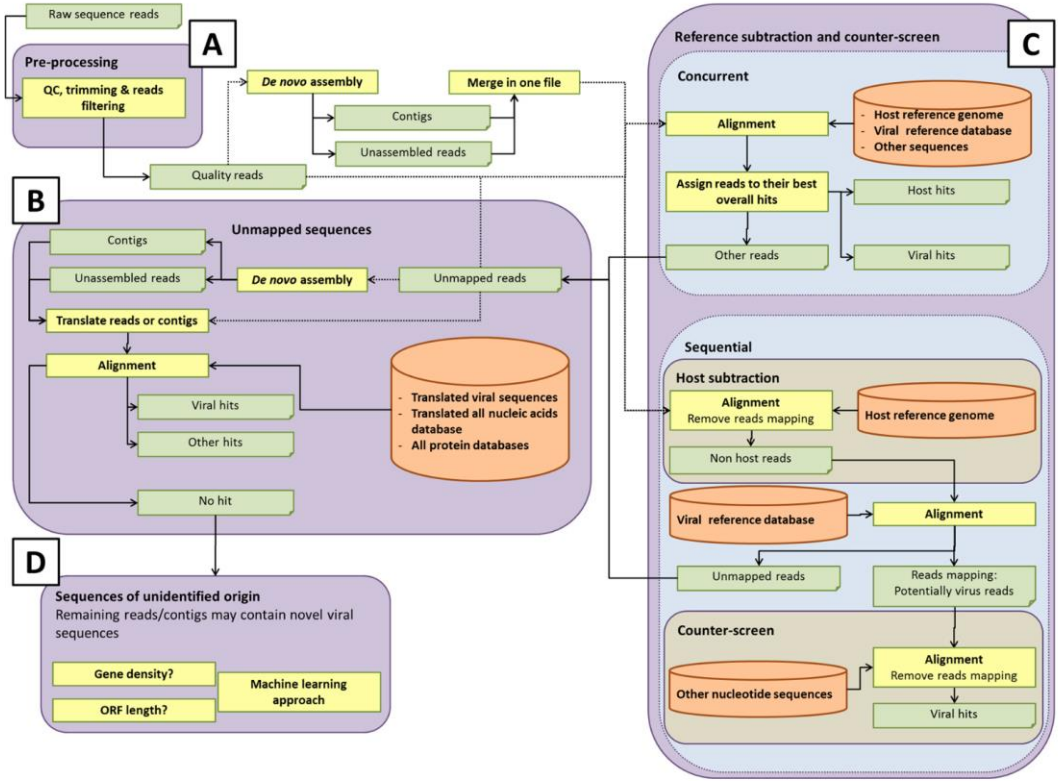


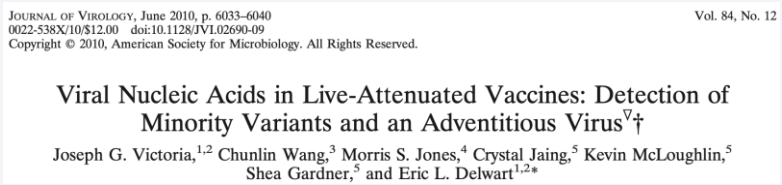
Figure 2. Potential pipelines for HTS data analysis for virus detection. Any given pipeline might use one or a combination of such paths, or others. See text for details. (**A**) Pre-processing, (**B**) Unmapped sequences, (**C**) Reference subtraction and counter-screen, and (**D**) Sequences of unidentified origin.

sanofi

PhyloID™, an evolving concrete instance within this design space

Purpose of PhyloID:
To identify putative adventitious agents (with a focus on viruses) in a sample (cell bank, seed lot, viral harvest, etc.) from a dataset of HTS-generated nucleotide sequences

Development of PhyloID started in 2010, as our response to the "circovirus crisis" (Victoria et al. 2010). Our assay has been GMP-validated since 2017.



JOURNAL OF VIROLOGY, June 2010, p. 6033–6040 Vol. 84, No. 12
0022-538X/10/$12.00 doi:10.1128/JVI.02690-09

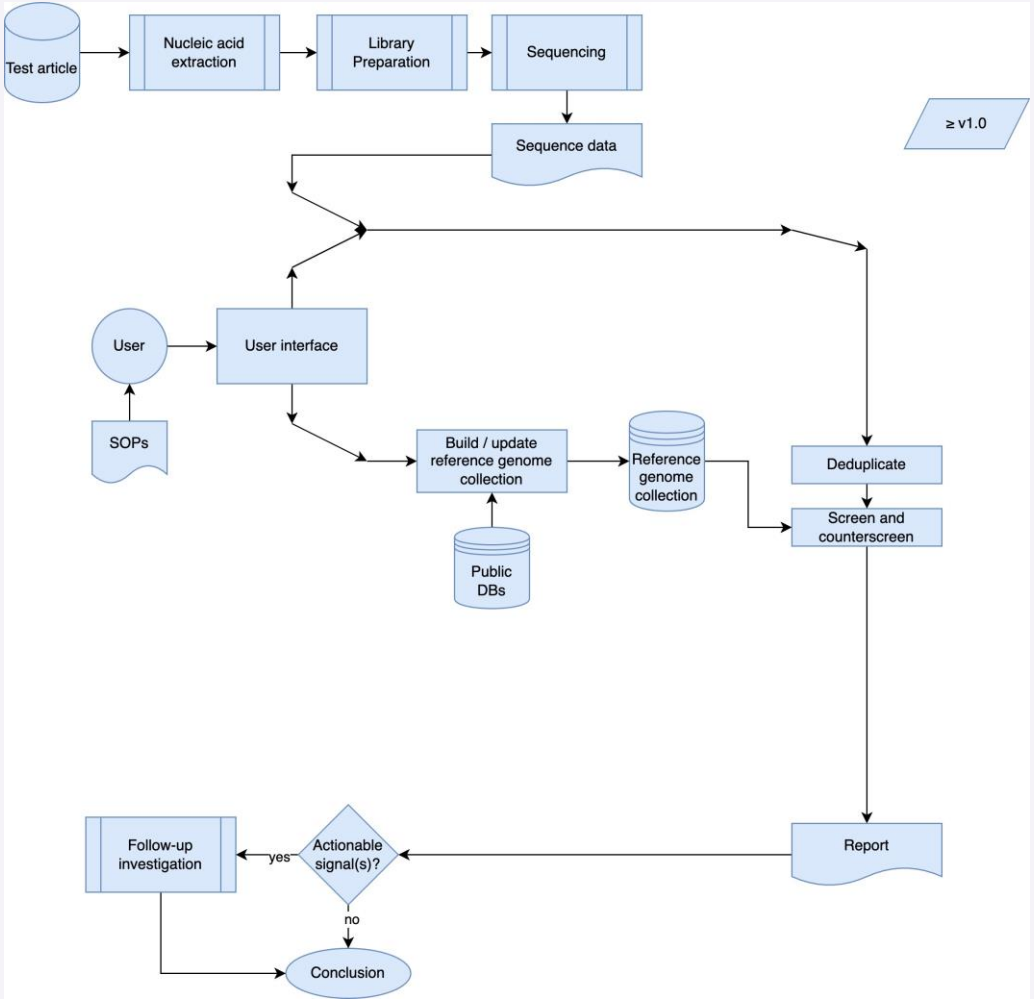
Viral Nucleic Acids in Live-Attenuated Vaccines: Detection of Minority Variants and an Adventitious Virus[▽]†

Joseph G. Victoria,[1,2] Chunlin Wang,[3] Morris S. Jones,[4] Crystal Jaing,[5] Kevin McLoughlin,[5] Shea Gardner,[5] and Eric L. Delwart[1,2]*

PhyloID™ v1

Modular to facilitate upgrades

- **Interface**: Command-line interface
-
-
-
- **Deduplicate**: *de novo* assembly; FASTQ to FASTA
- **Screen & counterscreen**: blastn-based on curated DB (RefSeq)
-
-
-
-
-



sanofi

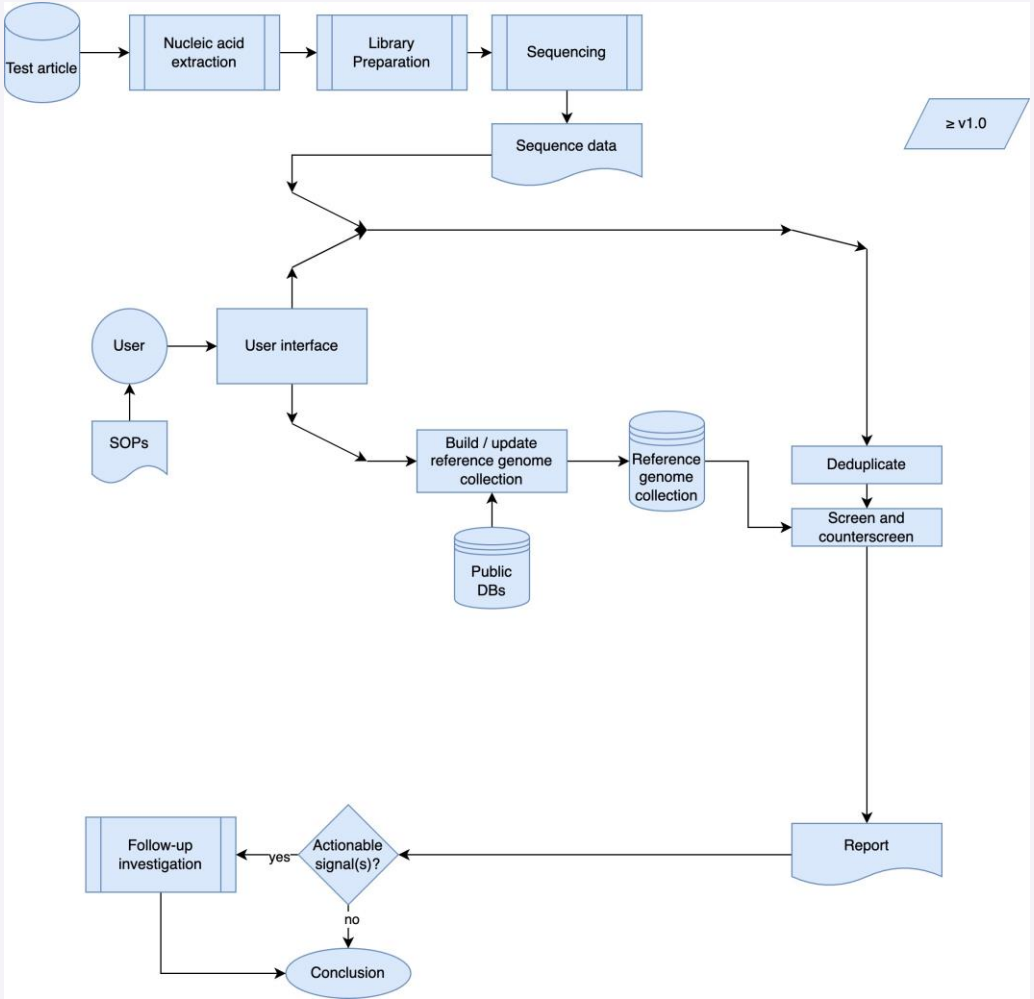
PhyloID™ v1 (2017-2022)

Cataloguing the Taxonomic Origins of Sequences from a Heterogeneous Sample Using Phylogenomics: Applications in Adventitious Agent Detection

Robert L. Charlebois, Siemon H. S. Ng, Lucy Gisonni-Lex, et al.

PDA J Pharm Sci and Tech **2014**, *68* 602-618

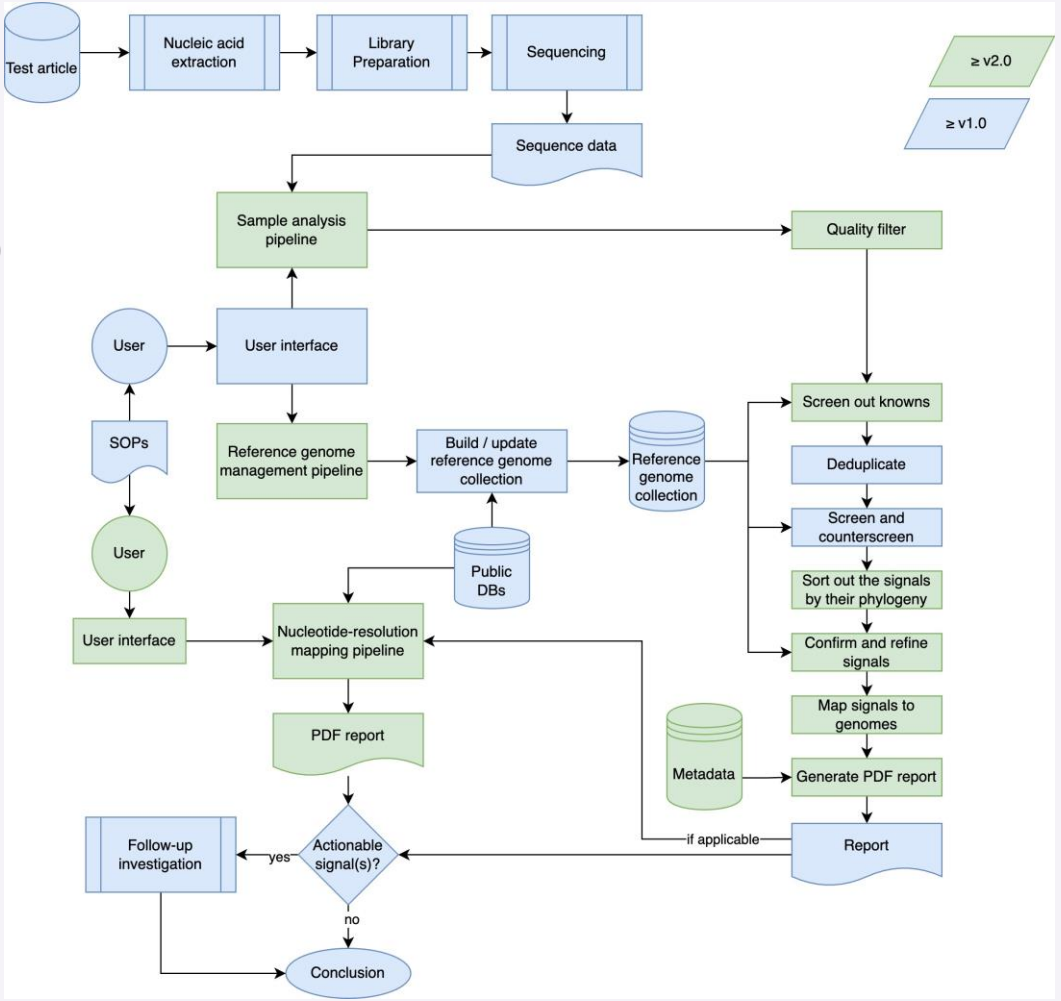
- Core strategy (Charlebois et al. 2014):
  - Build phylogenomic distance matrices from a trusted sequence repository
  - Reduce size of data for computational efficiency
  - Use a phylogenomic approach to identify sequences
- Benefits:
  - Good signal-to-noise: mitigate fears of opening Pandora's box
- Challenges:
  - Building the reference genome collection was slow and laborious
  - *De novo* assembly can lead to chimeric sequences and misidentification
  - Analysis pipeline was slow
  - Signals needed to be followed up manually



PhyloID™ v2.0

Modular to facilitate upgrades

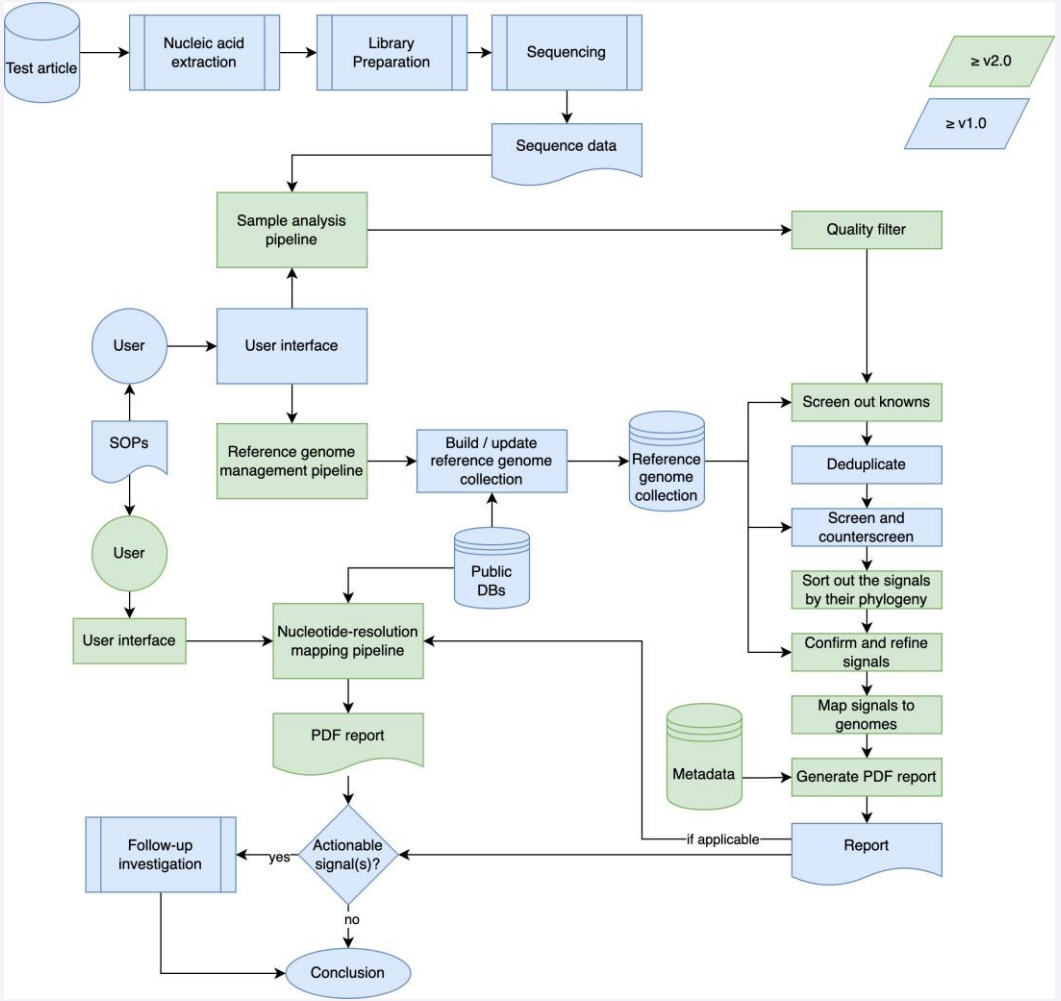
- Interface: web form on cloud
- **Quality filter**: mean composite Phred score ≥Q20
-
- **Screen out knowns**: high stringency; virus-masked or unmasked host
- Deduplicate: deduplication; FASTQ to FASTA
- Screen & counterscreen: blastn-based on curated DB (RefSeq)
- **Phylogeny**: signals in context on a tree
- **Confirm & refine**: at strain level using NCBI nt; protein matches using RVDB
- **Map signals**: meeting decision-tree criteria
- **Generate PDF**: automated, LaTeX-based
- **Fine mapping**: against the identified reference



bold = new
purple = improvements

PhyloID™ v2.0 (2022-2025)

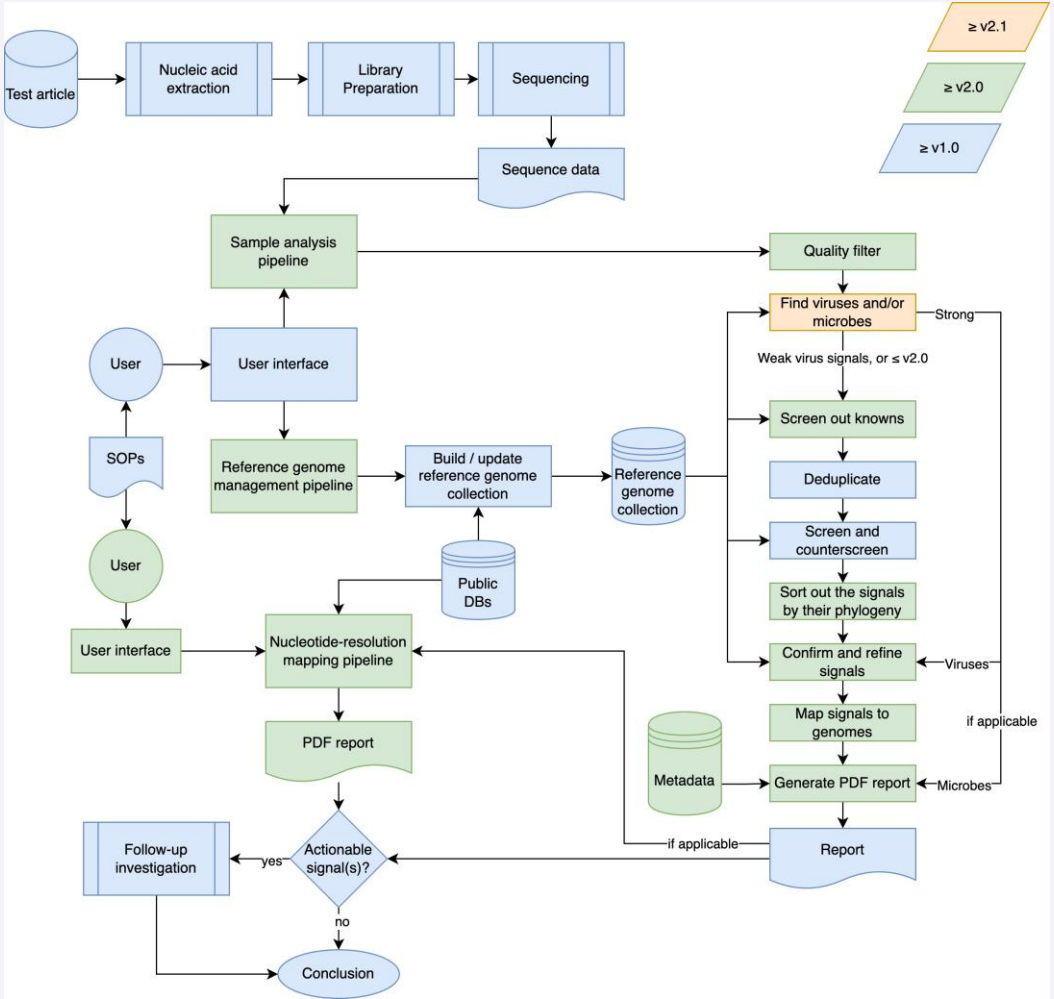
- Strategy improvements:
  - Cloud implementation: parallel processing
  - Quality filter avoids misclassification
  - Host filtering cleanly subtracts host, reducing size of dataset
  - Ordinary deduplication avoids chimeric sequences
  - K-mer based phylogenomics: reference curation automated and 1000× faster
  - Automated follow-up of signals
  - Semi-automated decision tree to find actionable signals
  - Better user interface; automated reporting
- Benefits:
  - Faster and more automated
- Challenges:
  - Still slow and expensive for bacteria
  - Incomplete support for retrovirus detection, necessitating e.g. F-PERT



PhyloID™ v2.1

Modular to facilitate upgrades

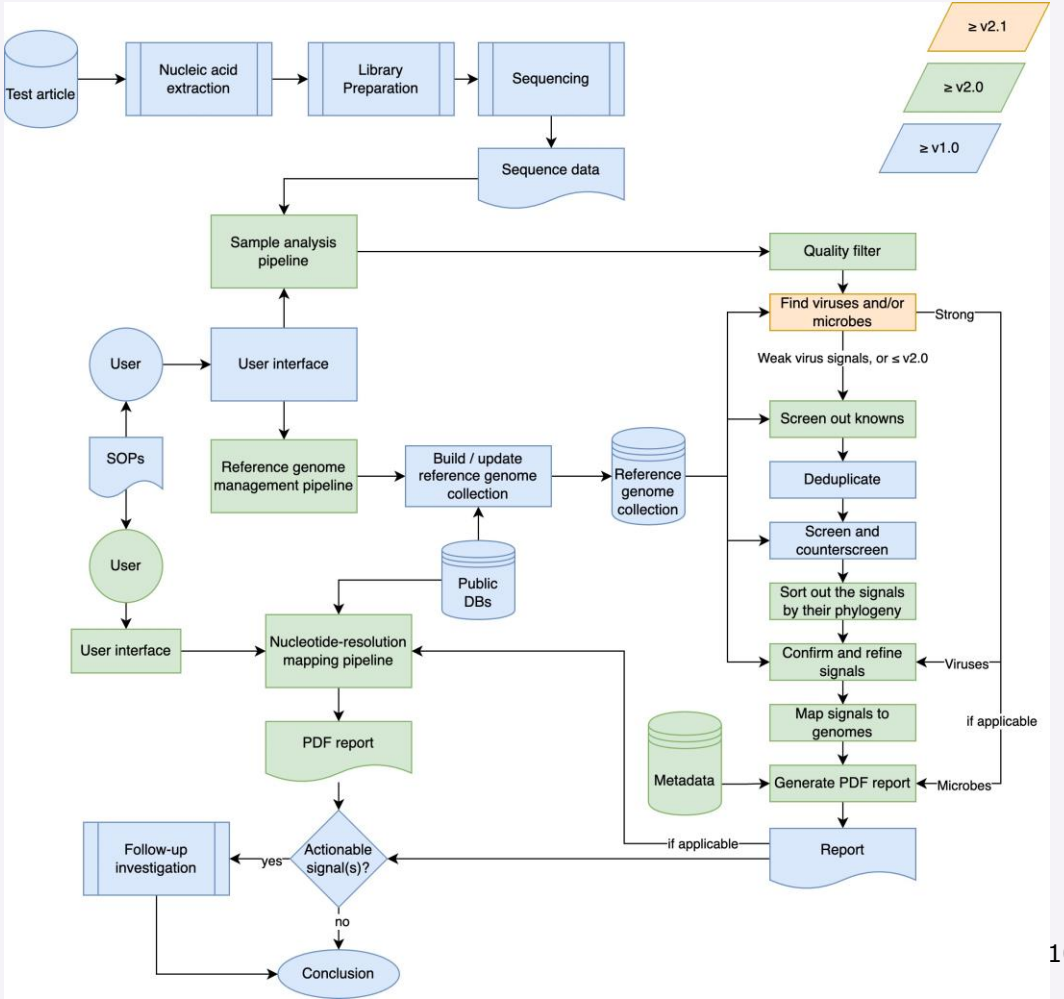
- Interface: web form on cloud
- Quality filter: mean composite Phred score ≥Q20
- **Find viruses/microbes**: position-aware k-mers
- Screen out knowns: high stringency; virus-masked or unmasked host
- Deduplicate: deduplication; FASTQ to FASTA
- Screen & counterscreen: blastn-based on curated DB (RefSeq+RVDB)
- Phylogeny: signals in context on a tree
- Confirm & refine: at strain level using NCBI nt
- Map signals: meeting decision-tree criteria
- Generate PDF: automated, JSON & LaTeX-based
- Fine mapping: against the identified reference



bold = new
purple = improvements

PhyloID™ v2.1 (2025–)

- Strategy improvements:
  - Use of a position-aware k-mer approach to find viruses and/or bacteria: sensitive, specific and fast
  - Continued use of v2.0 pipeline to resolve non-obvious viral signals
  - More complete virus database: bringing retrovirus detection into scope
  - More automation for the decision tree
  - JSON files for better interoperability
  - Simpler interface; nicer reports
- Benefits:
  - Faster, more affordable, more automated
  - Bacteria and retroviruses now in scope
- Challenges:
  - Need better models for background signals: work in progress
  - Incomplete support for truly novel viruses; tracking the literature for insights



Summary

- First: getting the job done to meet our viral safety commitments (v1)
  - Mitigate the risks of *missing something*
    - extra caution = false positives ⇀ for SME scrutiny
  - Mitigate the risks of *finding something*
    - extensive development prior to validation informing the design of decision trees
    - deep discussions (internal, and within AVDT{U/I}G), and stakeholder education
- Next: getting the job done efficiently to better meet project timelines (v2.0)
  - v1 was not scalable
    - its internal acceptance and adoption meant greater demand
    - its need for hands-on hard-core bioinformatics meant greater specialist workload
  - Leverage experience to understand what impacts risk
    - for both false negatives and false positives, informing the design of new application modules
    - automating decision trees ⇀ making the software the front-line SME
- Then: expanding its scope and setting up for the future (v2.1)
  - Addressing recognized gaps, in consideration of future testing package streamlining
    - retroviruses, microorganisms
  - Further performance improvements
  - Setting up data structures for later PhyloID versions, to better manage background signals

Acknowledgments

- At Sanofi:
  - Artur Pedyczak
  - Song Sun
  - Alejandra Chavez Carbajal
  - Carine Logvinoff
  - Lucy Gisonni-Lex
  - Lauren Rodrigues
  - Jacek Remani
  - Shanaz Gilchrist
  - Daniel Biehle
  - Tuan Nguyen
- Formerly at Sanofi:
  - Siemon Ng
  - Sarmitha Sathiamoorthy
  - Laurent Mallet
  - Walter Ungureanu
- Outside Sanofi:
  - AVDTWG (formerly AVDTIG), especially subgroup D/E

sanofi

Thank you

Note: RL Charlebois is a Sanofi employee and may hold shares and/or stock options in the company.

sanofi