Leiden University
Medical Center

ENNGS benchmark of thirteen bioinformatic pipelines for metagenomic virus diagnostics using datasets from clinical samples

ICCMg 2021-10-21

Jutte de Vries
LUMC, Dept. Medical Microbiology,
Leiden, the Netherlands
ON BEHALF OF THE ENNGS

ENNGS
ESCV NETWORK ON NEXT-GENERATION SEQUENCING

Introduction viral metagenomic sequencing (mNGS)

- Viral mNGS is increasingly being used in virology laboratories for difficult to diagnose cases
- The current main clinical application is encephalitis of unknown cause, but considered useful in a growing number of other clinical syndromes
- The performance of mNGS is largely dependent on accurate bioinformatic analysis, on both classification algorithms and databases

Challenges bioinformatic analysis in the diagnostic lab

- A wide range of metagenomic pipelines and taxonomic classifiers have been developed but commonly for the purpose of biodiversity/microbiome studies
- Potential false-negative and false-positive bioinformatic classification results can have significant consequences for patient care
- Most reports on bioinformatic tools for metagenomic analysis for virus diagnostics typically describe algorithms and validations of single in-house pipelines developed by the authors themselves

Aim

To conduct a benchmark of bioinformatic pipelines using viral metagenomic datasets derived from clinical samples, in order to assist laboratories with selection and optimization of tools to be implemented for clinical use

ESCV Network on NGS (ENNGS)

Established in 2018 under the auspices of the European Society for Clinical Virology
Participants from >15 countries: UK, IR, GE, NO, SW, FI, DK, AU, FR, ES, IT, IS, GR, CZ, TU, BE, NL

AIMS

- to bring together professionals involved in viral diagnostics using NGS
- develop, improve and standardize viral NGS diagnostics
- sharing data, experiences, methodologies; METASHARE platform veb.lumc.nl/CliniMG/metashare.cgi

ESCV

European Society for Clinical Virology

ENNGS

ESCV NETWORK ON NEXT-GENERATION SEQUENCING

Journal of Clinical Virology 134 (2021) 104691

Contents lists available at ScienceDirect

Journal of Clinical Virology

journal homepage: www.elsevier.com/locate/jcv

Recommendations for the introduction of metagenomic high-throughput sequencing in clinical virology, part I: Wet lab procedure

F. Xavier López-Labrador [a,b], Julianne R. Brown [c], Nicole Fischer [d], Heli Harvala [e], Sander Van Boheemen [f], Ondrej Cinek [g], Arzu Sayiner [h], Tina Vasehus Madsen [i], Eeva Auvinen [j], Verena Kufner [k], Michael Huber [k], Christophe Rodriguez [l], Marcel Jonges [m,n], Mario Hönemann [o], Petri Susi [p], Hugo Sousa [q,r,s,t], Paul E. Klapper [u], Alba Pérez-Cataluña [v], Marta Hernandez [w], Richard Molenkamp [f], Lia van der Hoek [m,n], Rob Schuurman [x], Natacha Couto [y,z], Karoline Leuzinger [A,B], Peter Simmonds [C], Martin Beer [D], Dirk Höper [D], Sergio Kamminga [E], Mariet C.W. Feltkamp [E], Jesús Rodríguez-Díaz [F], Els Keyaerts [G], Xiaohui Chen Nielsen [i], Elisabeth Puchhammer-Stöckl [H], Aloys C.M. Kroes [E], Javier Buesa [F], Judy Breuer [c], Eric C. J. Claas [E], Jutte J.C. de Vries [E,*], on behalf of the ESCV Network on Next-Generation Sequencing

Journal of Clinical Virology 138 (2021) 104812

Contents lists available at ScienceDirect

Journal of Clinical Virology

journal homepage: www.elsevier.com/locate/jcv

Recommendations for the introduction of metagenomic next-generation sequencing in clinical virology, part II: bioinformatic analysis and reporting

Jutte J.C. de Vries [a,*], Julianne R. Brown [b], Natacha Couto [c], Martin Beer [d], Philippe Le Mercier [e], Igor Sidorov [a], Anna Papa [f], Nicole Fischer [g], Bas B. Oude Munnink [h], Christophe Rodriquez [i], Maryam Zaheri [j], Arzu Sayiner [k], Mario Hönemann [l], Alba Perez Cataluna [m], Ellen C. Carbo [a], Claudia Bachofen [n], Jakub Kubacki [n], Dennis Schmitz [o], Katerina Tsioka [f], Sébastien Matamoros [p], Dirk Höper [d], Marta Hernandez [q], Elisabeth Puchhammer-Stöckl [r], Aitana Lebrand [e], Michael Huber [j], Peter Simmonds [s], Eric C.J. Claas [a], F. Xavier López-Labrador [t,u,v,**], on behalf of the ESCV Network on Next-Generation Sequencing

Methods: datasets

13 clinical metagenomic datasets (FASTQ) from samples well-characterized by PCR from patients with encephalitis or respiratory complaints

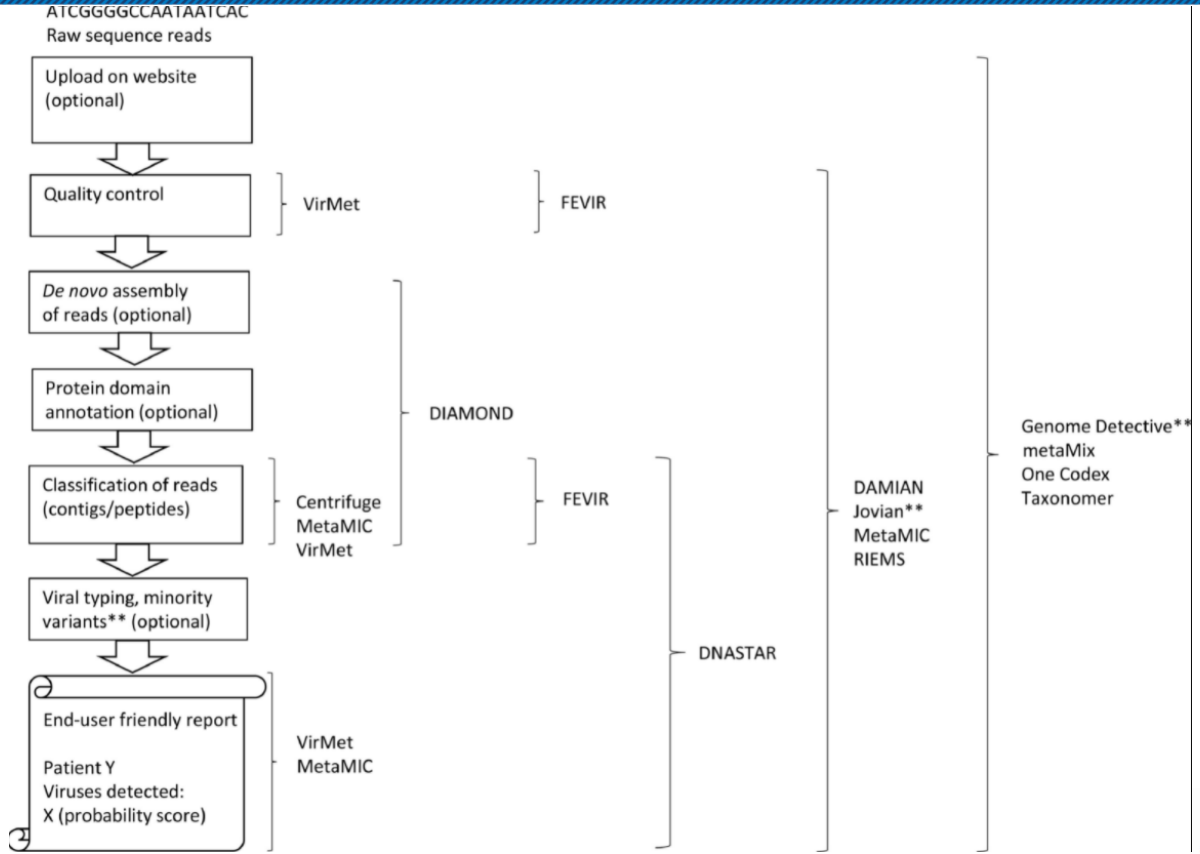
- CSF (n=4)
- Brain biopsies (n=3)
- Nasopharyngeal swabs (n=3)
- Nasal washings (n=1)
- Bronchoalveolar lavage (n=1)
- Plasma (n=1)

Methods: sequencing

Site 1: mRNA seq, Illumina's TruSeq Stranded mRNA LT prep kit, NextSeq500
Site 2: RNA/DNA seq, NEBNext Ultra Directional RNA prep kit with in-house adaptations, NextSeq500/NovaSeq6000

- Datasets were analysed in a blinded fashion by participants
- Qualitative and quantitative performance, PCR as gold standard
- Parameters: virus pathogen detection, taxonomic classification level, target read count, horizontal genome coverage, computational time, user-friendliness and output formats

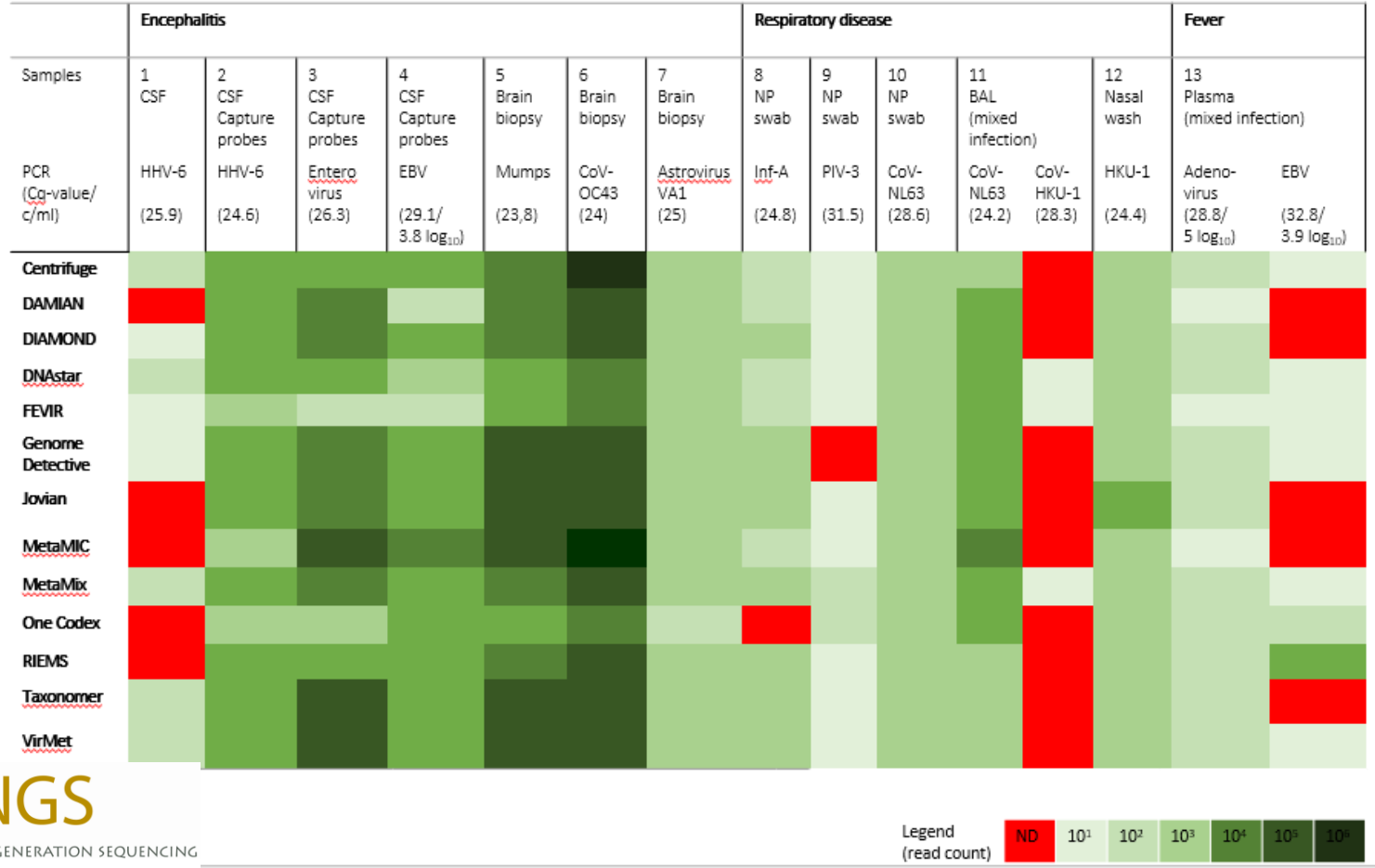
Bioinformatic workflows; 13 pipelines



Short summary of overall pipeline characteristics

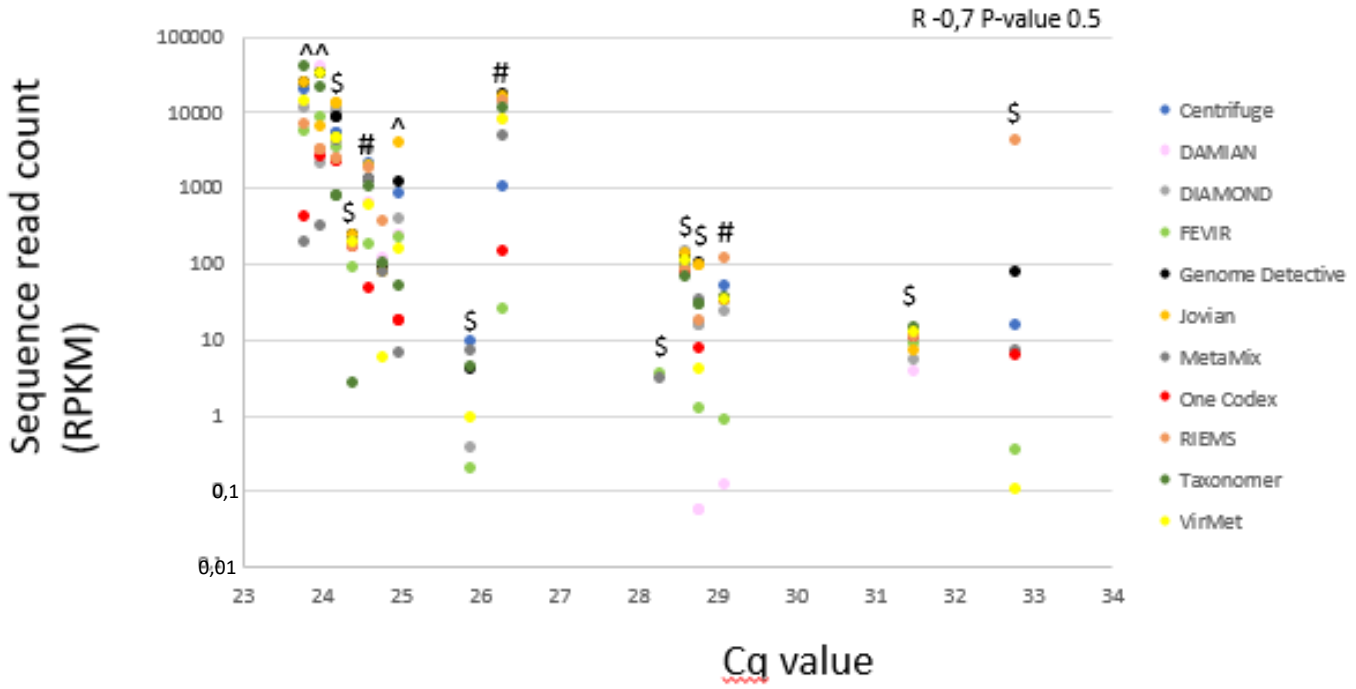
- Nine pipelines were implemented in patient care, 3 of them accredited: MetaMIC, metaMix, VirMEt
- Majority developed or adapted the pipeline at a local site
- Four pipelines are commercially available and web-based:
  - DNASTAR
  - GenomeDetective
  - One Codex
  - Taxonomer
- Publicly available: Centrifuge, DAMIAN
- (Adapted versions of) databases NCBI's Genbank nt and RefSeq were most commonly used
- *De novo* assembly was part of 6 out of 13 pipelines
- Classification was based on nt similarity (8/13), AA similarity (2/13) or a combination of both (3/13)

Qualitative results, overall sensitivity 80-100%



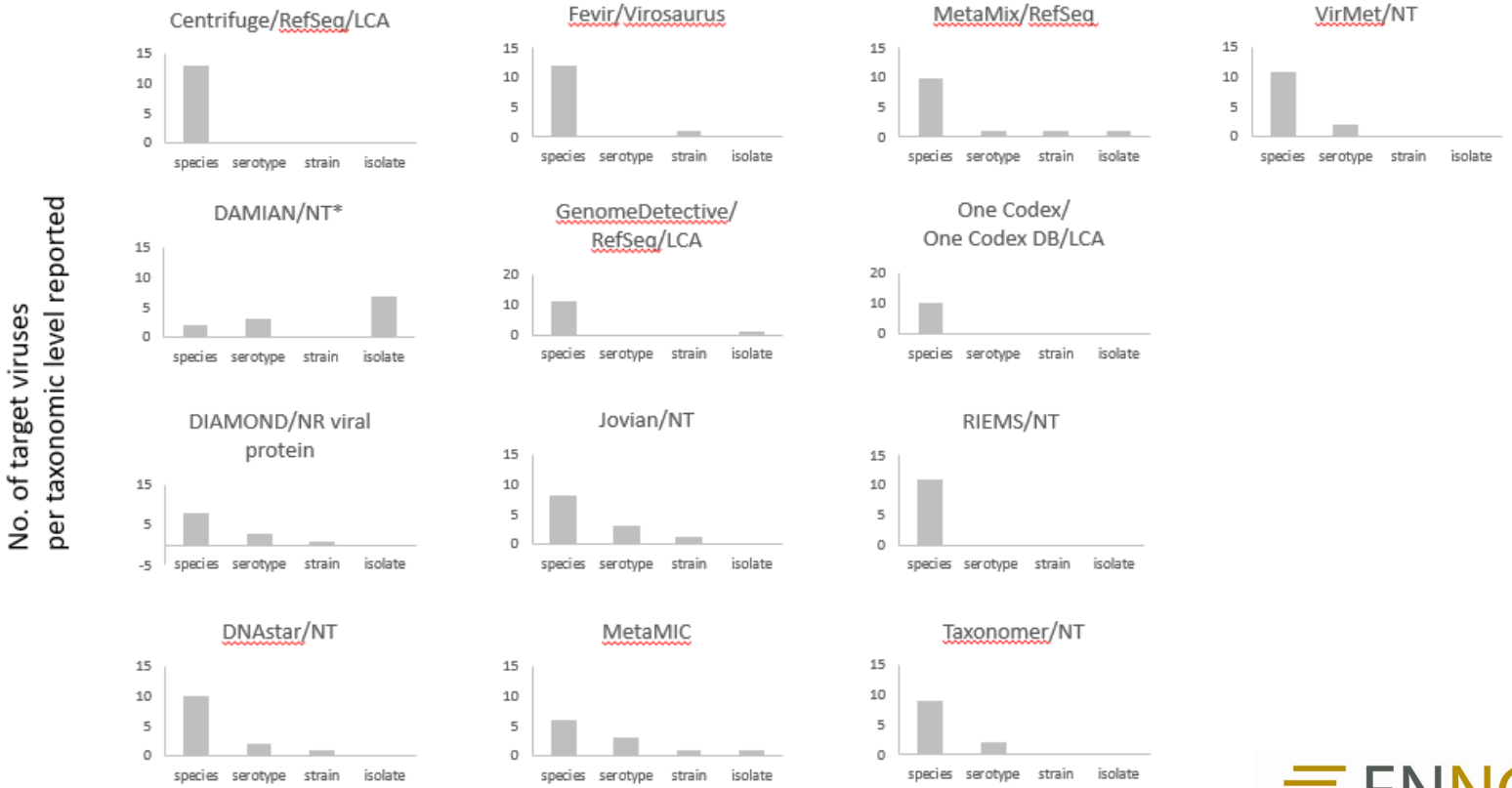
ENNGS
ESCV NETWORK ON NEXT-GENERATION SEQUENCING

Semi-quanitative results, sensitivity



(^ mRNA sequencing, $ RNA/DNA sequencing, and #: a captured approach using probes targeting vertebrate viruses

Classification level



Additional virus hits

Either not tested for by RT-PCR or RT-PCR negative, were reported by 11 out of 13 pipelines, and in one or more samples

Reported by multiple pipelines and absent in the negative run control (not available for the participants):

Not tested for by RT-PCR

- human retrovirus RD114* (2-2102 reads, up to 28% genome coverage)
- feline leukemia virus* (2-1406 reads)
- torque-teno virus (TTV)* (18-66 reads, up to 7% genome coverage)
- polyomaviruses* (5-41 reads, up to 37% genome coverage)
- bovine viral diarrhea virus (BVDV) (6-220 reads, likely FBS contaminants)
- dengue virus (18-370 reads)

* Given their association with the host (integrated or commensal) likely true positive findings

Positive predictive value

When considering viral mNGS hits with negative RT-PCR results:

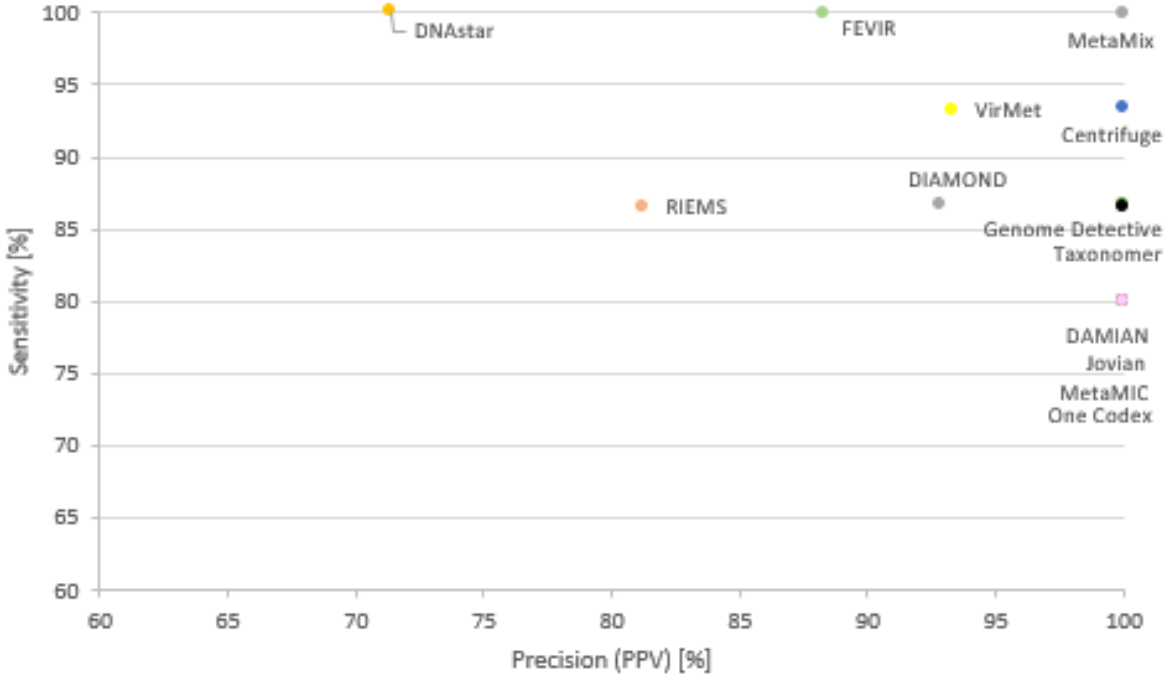
CoV-NL63 (1 read), PIV-4 (2-6 reads), HRV-C (2-4 reads), CoV-OC43 (5 reads), INF-B (2 reads)

> PPV 71-100%

No distinction could be made between assignments of sequences genuinely present e.g. by index hopping (which was suspected given the low number of reads), false negative by PCR due to primers/probes mismatches, and false positive assignments

Overall score (sensitivity/PPV)

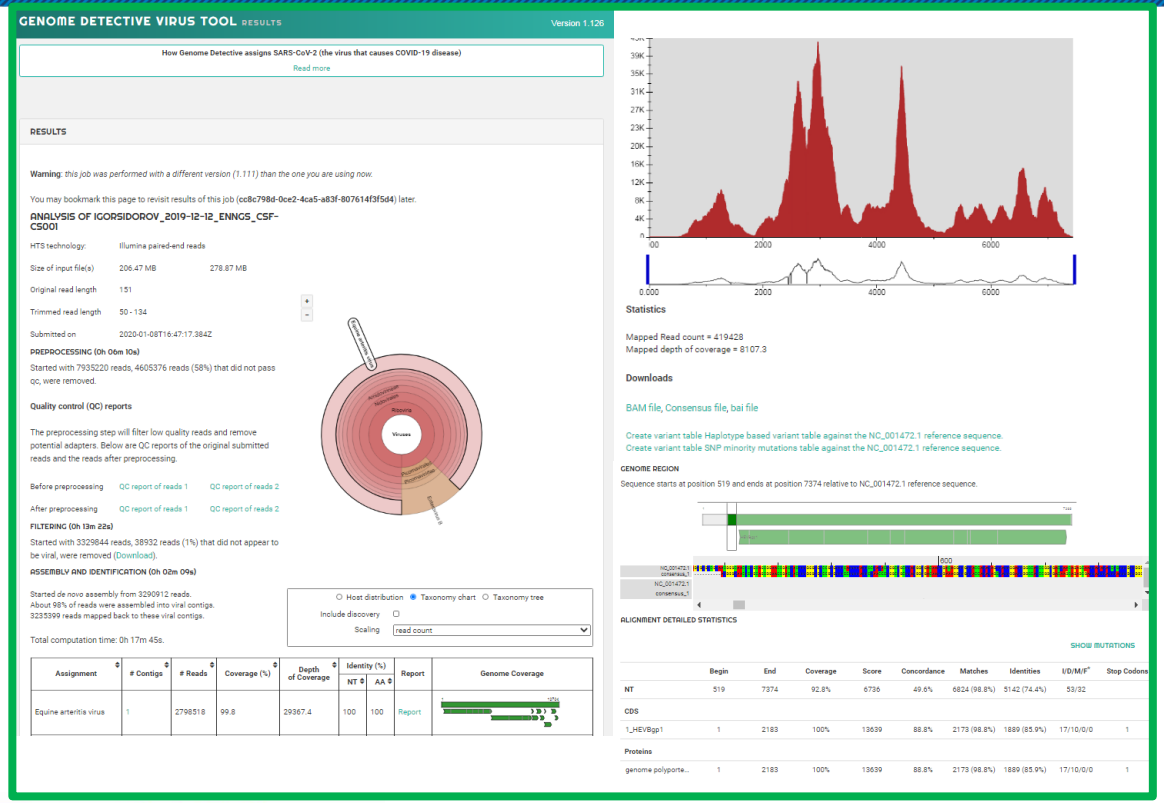
	PCR+	PCR -	
mNGS +	TP	FP	PPV
mNGS -	FN	TN?	NPV
	Sensitivity	Specificity	



Reporting criteria

- Parameters used for defining a positive result: read counts, horizontal genome coverage (some of the participants), post-probability scores (1/13), ROC curve (1/13)
- BLAST analysis of matching sequences was commonly used to exclude misassignments/ to confirm true positive hits
- Confirmatory PCR (outside this benchmarking) before reporting, one participant indicated that this was not needed based on their validation studies

User-friendly output formats



metaMix hosted by Bluebee

RNA-Seq Encephalitis Diagnostics

Pipeline Run Details

User Reference:	GOSHmeta3	Pipeline:	GOSH RNA-Seq Encephalitis Diagnostics 1.2.0
Request Date:	Sep, 10 2019 10:58:33	Start Date:	Sep, 10 2019 11:00:23
Duration:	14h 58m 33s	Requestor:	Dr. Julianne Brown
User Tags:			

Input Data

UCLGNS1212-13M1974-B_S7_R1_001.fastq.gz

File Name:	UCLGNS1212-13M1974-B_S7_R1_001.fastq.gz	File Path:	UCLGNS1212-13M1974-B_S7_R1_001.fastq.gz
Size:	5.57 GB	Format:	FASTQ
Creation Date:	Sep, 10 2019 08:55:02	User Tags:	
Run In Tags:	GOSHmeta3	Connector Tags:	Upload

UCLGNS1212-13M1974-B_S7_R2_001.fastq.gz

File Name:	UCLGNS1212-13M1974-B_S7_R2_001.fastq.gz	File Path:	UCLGNS1212-13M1974-B_S7_R2_001.fastq.gz
Size:	5.7 GB	Format:	FASTQ
Creation Date:	Sep, 10 2019 08:47:18	User Tags:	
Run In Tags:	GOSHmeta3	Connector Tags:	Upload

Results

	"taxonID"	"scientName"	"finalAssignments"	"poster.prob"	"log10BF"
"8"	"unknown"	"unknown"	30490	1	NA
"7"	"9606"	"Homo sapiens"	28586	1	28977.6477200774
"6"	"645687"	"Astrovirus VA1"	2423	1	9562.99329606601
"1"	"10090"	"Mus musculus"	536	1	684.019570605247
"2"	"28090"	"Acinetobacter lwoffii"	25	1	135.6328430578
"4"	"469"	"Acinetobacter"	19	0.99	57.62766626128
"3"	"43675"	"Rothia mucilaginosa"	14	1	109.876588922052
"5"	"488"	"Neisseria mucosa"	11	0.94	14.9840642137569

List of detected species (presentSpecies_assignedReads.tsv)

Conclusions

- First large-scale international benchmarking study using datasets from clinical samples and pipelines currently applied in a large series of clinical viral diagnostic laboratories
- All of the participants used different classification tools, though no selection of laboratories using different tools was made in advance
- Overall high sensitivity for detecting viral pathogens with relatively high viral loads (Cq-values <28)
- Lower abundant pathogens and mixed infections were only detected by 3/13 the pipelines
- Overall sensitivity 80-100%, PPV 71-100%
- No clear differences were observed in terms of performance based on nucleotide-based classification versus amino acid-based classification and *de novo* assembly-based algorithms versus read based classification.

Discussion

Reported read counts and genome coverage varied between pipelines up to several orders of magnitude
Differences observed in limits of detection for samples with low viral loads

- Differences in reporting of unique versus non-uniquely mapped sequences may be underlying

PPV calculations were hampered by the intrinsic inability to distinguish between sequences actually present in the dataset that might be undetected by RT-PCR (index hopping, primer mismatches, prep contaminants)

Given the inclusion of commercially available pipelines with fixed databases, it was not feasible to compare the different tools with one standardised database at the local sites, but the design did allow for comparison of the complete pipeline in use for clinical diagnostics, from QC to reporting algorithms including posterior probability scores

No conclusions can be drawn on the limit of detection of the full metagenomic workflows used in each specific laboratorie since this is dependent on the wet lab procedure, sequencer, and specific cut-of/prob. values

Leiden University Medical Center

Acknowledgements

Julianne R. Brown, Sofia Morfopoulou and Judith Breuer (UK, London, GOSH)

Nicole Fischer and Jiabin Huang (GE, Hamburg, UMCH-E)

Igor A. Sidorov, Eric C.J. Claas and Aloys Kroes (NL, Leiden, LUMC)

Bas B. Oude Munnink and Sander van Boheemen (NL, Rotterdam, EMC)

Arzu Sayiner and Alihan Bulgurcu (TU, Izmir, DEU)

Christophe Rodriguez and Guillaume Gricourt (FR, Paris, HHM)

Els Keyaerts and Leen Beller (BE, Leuven, KUL)

Claudia Bachofen and Jakub Kubacki (SW, Zurich, UZ)

Samuel Cordey and Florian Laubscher (SW, Geneva, UHG)

Dennis Schmitz (NL, Bilthoven, RIVM)

Martin Beer and Dirk Hoeper (GE, Greifswald, FLI)

Michael Huber, Verena Kufner and Maryam Zaheri (SW, Zurich, UZ)

Aitana Lebrand (SW, Geneva, SIB)

Anna Papa (GR, Thessaloniki, AUT)

F. Xavier Lopez-Labrador (SP, Valencia, FISABIO)

ENNGS

ESCV NETWORK ON NEXT-GENERATION SEQUENCING