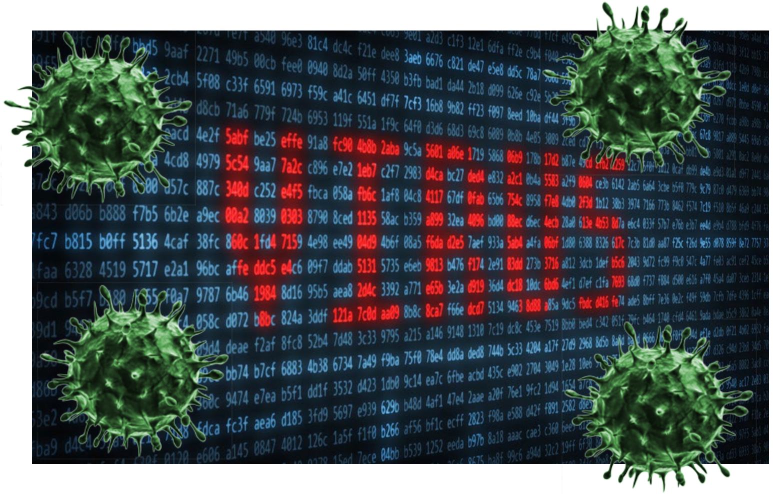
Leiden University
Medical Center

Coronavirus discovery by metagenomic sequencing: a tool for pandemic preparedness



Ellen Carbo
Clinical Microbiological Laboratory

ICCMG 15-10-2020

DOI: 10.1016/J.JCV.2020.104594

Pandemic Prepardness

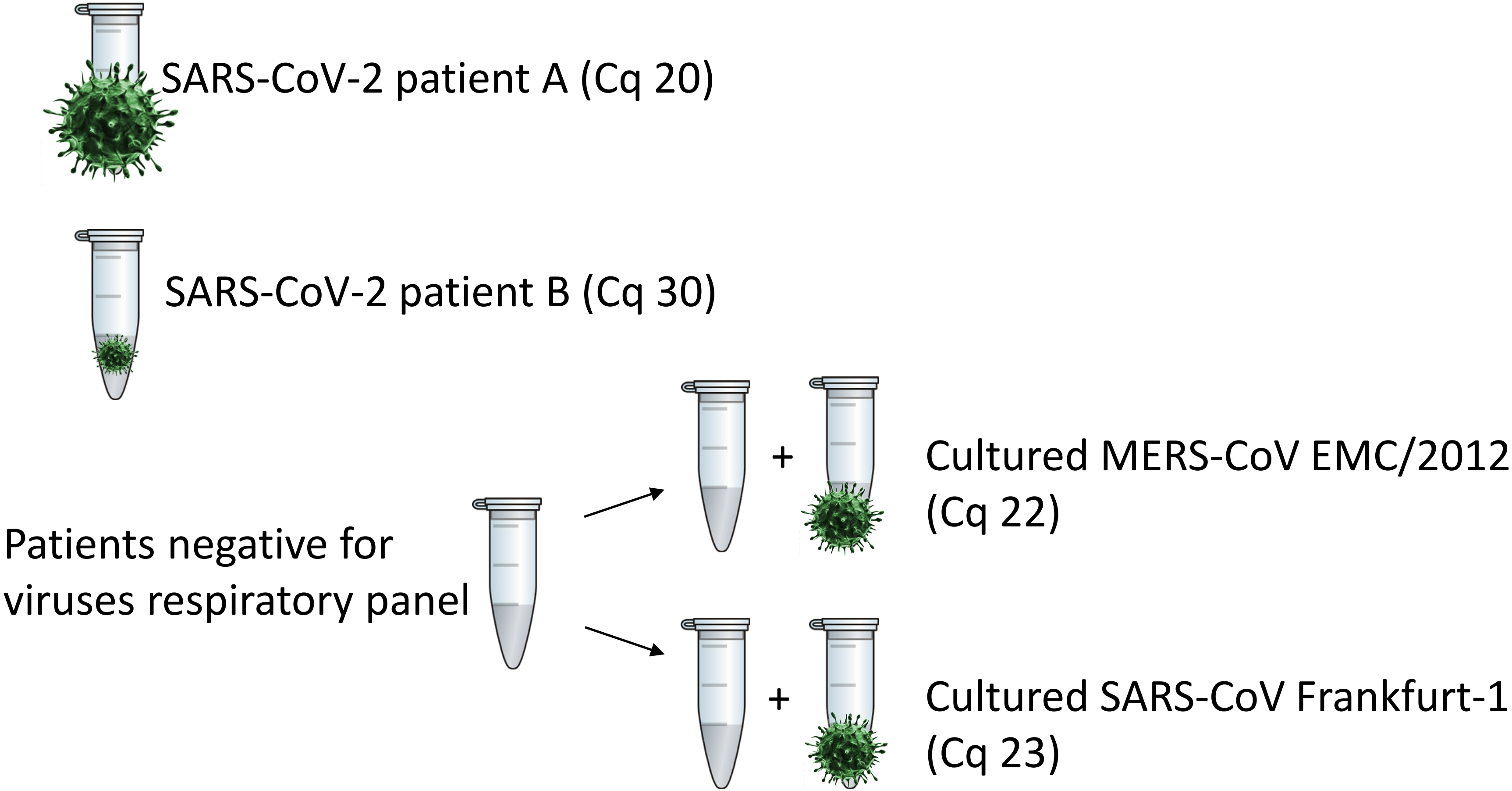
To detect and identify a novel virus as quickly as possible

Metagenomic classifiers use reference indexes with only known viruses
Not the unknown ones

In this study we validate the performance of a virus discovery model

doi: 10.1016/j.jcv.2020.104594

Samples



SARS-CoV-2 patient A (Cq 20)

SARS-CoV-2 patient B (Cq 30)

Patients negative for viruses respiratory panel

+

Cultured MERS-CoV EMC/2012 (Cq 22)

+

Cultured SARS-CoV Frankfurt-1 (Cq 23)

Methods

NA extraction - MP96 Roche

Confirm Cq value pathogen by qPCR

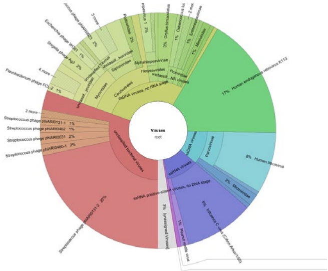
Library Prep - NEBNext® Ultra™ Directional RNA Library Prep Kit for Illumina®

(Virocap SeqCap EZ HyperCap for SARS-CoV-2 patients)

Sequencing – Novaseq Genomescan

Data analysis

- Trimming, fastqc and removal of host reads
- Centrifuge classification tool and Refseq database
- Genome Detective
- De novo assembly and blast



Build references

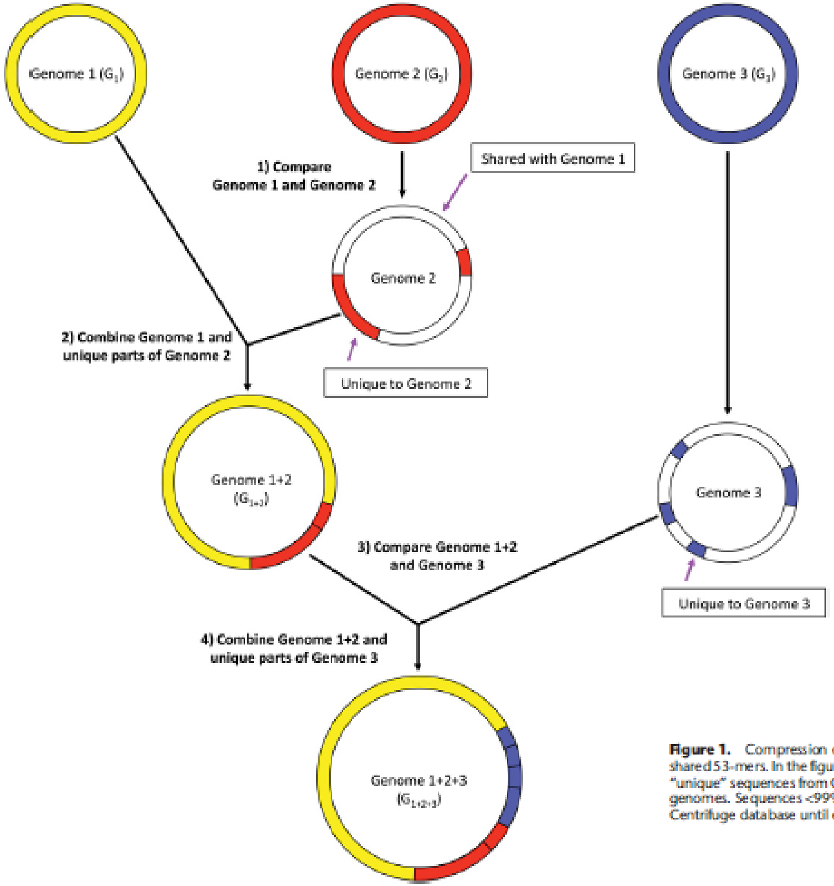
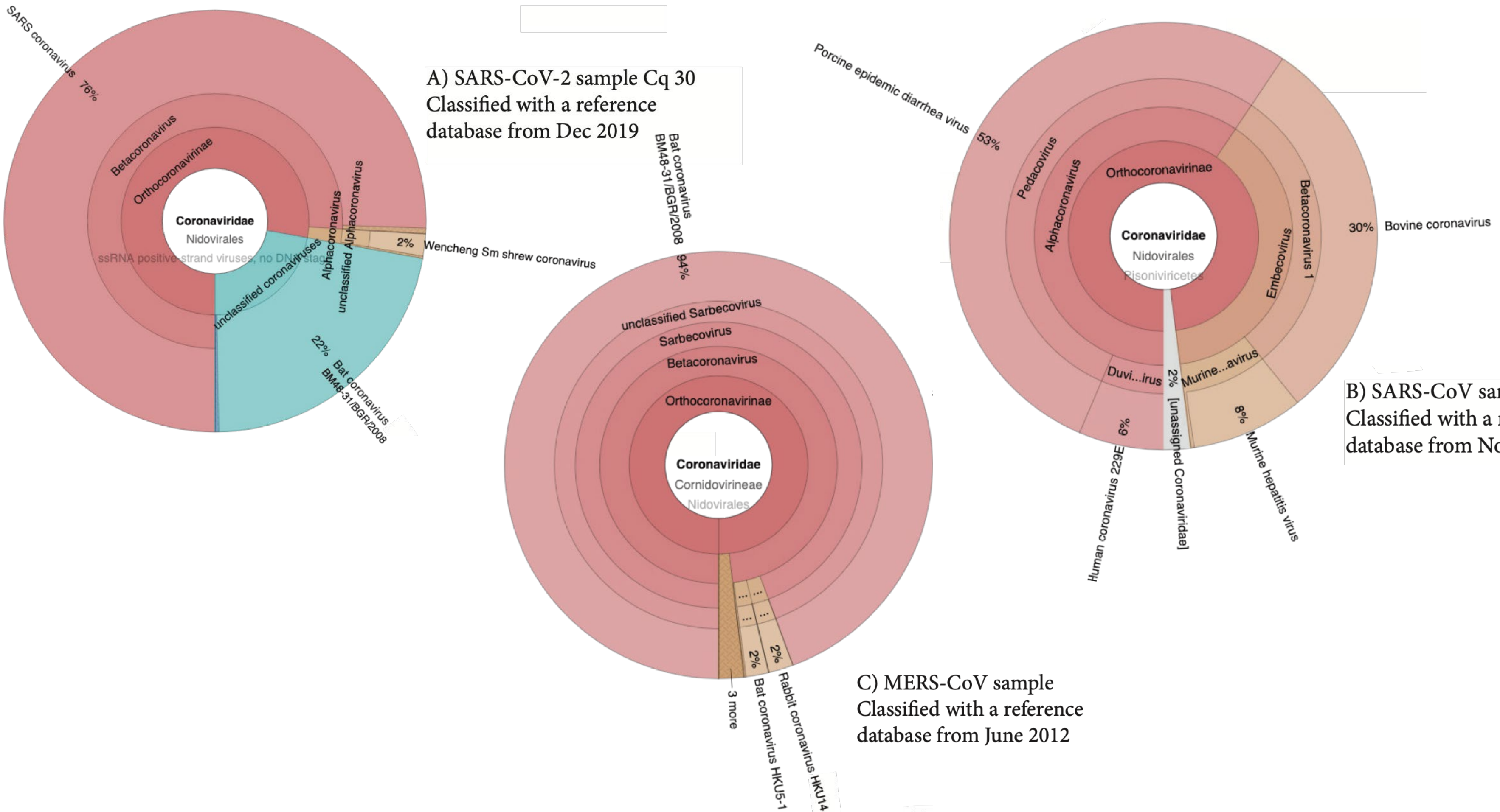

Genome 1 (G_1)
Genome 2 (G_2)
Genome 3 (G_3)
1) Compare Genome 1 and Genome 2
Shared with Genome 1
Genome 2
2) Combine Genome 1 and unique parts of Genome 2
Unique to Genome 2
Genome 1+2 (G_{1+2})
Genome 3
3) Compare Genome 1+2 and Genome 3
Unique to Genome 3
4) Combine Genome 1+2 and unique parts of Genome 3
Genome 1+2+3 (G_{1+2+3})

Figure 1. Compression of genome sequences before building the Centrifuge index. All genomes are compared and similarities are computed based on shared 53-mers. In the figure, genomes G_1 and G_2 are the most similar pair. Sequences of G_2 that are ≥99% identical to G_1 are discarded, and the remaining "unique" sequences from G_2 are added to genome G_1, creating a merged genome, G_{1+2}. Similarity between all genomes is recomputed using the merged genomes. Sequences <99% identical in genome G_3 are then added to the merged genome, creating genome G_{1+2+3}. This process repeats for the entire Centrifuge database until each merged genome has no sequences ≥99% identical to any other genome.

- One of Dec last year
- One with viruses of before SARS-CoV 2002
- One with viruses of before MERS-CoV 2012

Centrifuge classification results



A) SARS-CoV-2 sample Cq 30
Classified with a reference
database from Dec 2019

B) SARS-CoV sample
Classified with a reference
database from Nov 2002

C) MERS-CoV sample
Classified with a reference
database from June 2012

Centrifuge classification results (2)

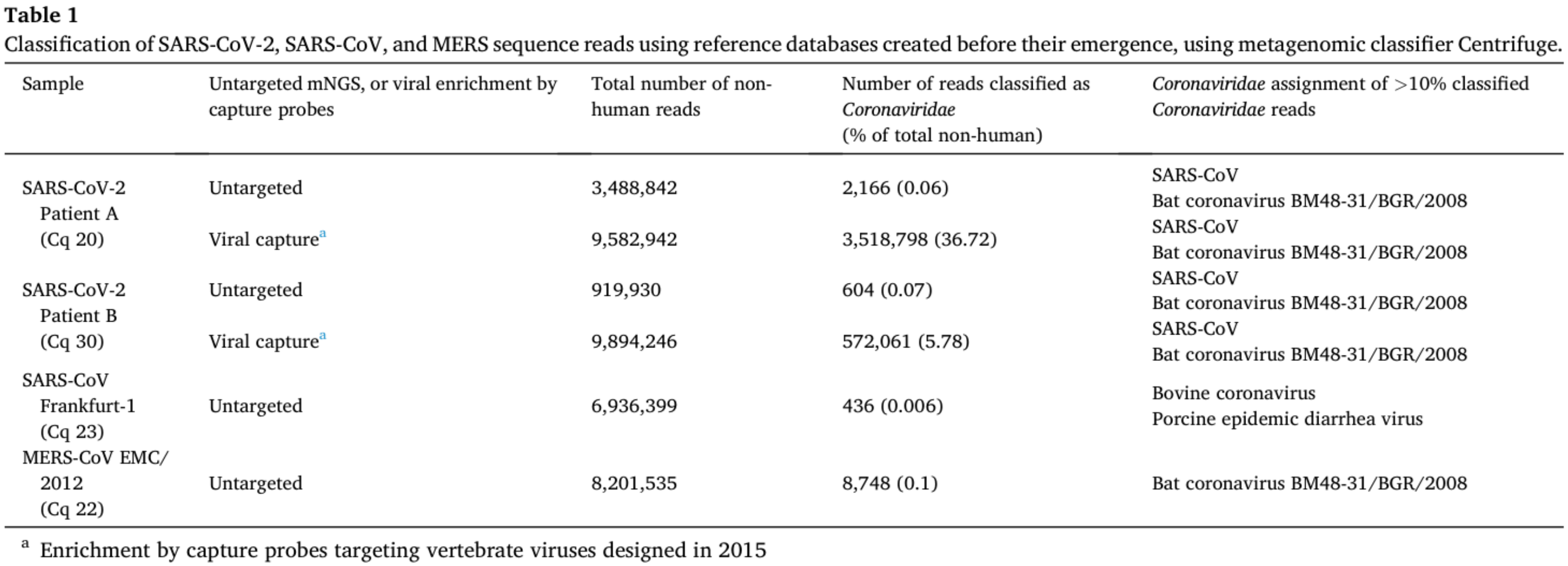
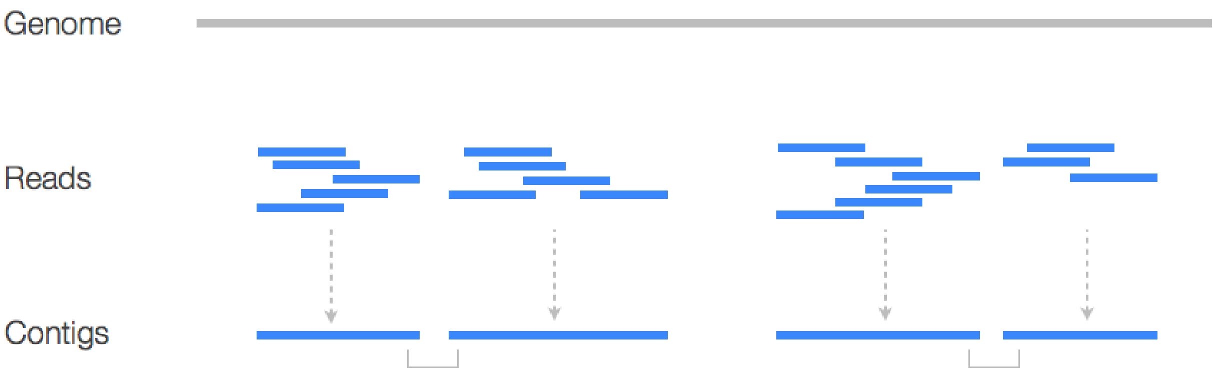


Table 1
Classification of SARS-CoV-2, SARS-CoV, and MERS sequence reads using reference databases created before their emergence, using metagenomic classifier Centrifuge.

Sample	Untargeted mNGS, or viral enrichment by capture probes	Total number of non-human reads	Number of reads classified as *Coronaviridae* (% of total non-human)	*Coronaviridae* assignment of >10% classified *Coronaviridae* reads
SARS-CoV-2 Patient A (Cq 20)	Untargeted	3,488,842	2,166 (0.06)	SARS-CoV Bat coronavirus BM48-31/BGR/2008
	Viral capture[a]	9,582,942	3,518,798 (36.72)	SARS-CoV Bat coronavirus BM48-31/BGR/2008
SARS-CoV-2 Patient B (Cq 30)	Untargeted	919,930	604 (0.07)	SARS-CoV Bat coronavirus BM48-31/BGR/2008
	Viral capture[a]	9,894,246	572,061 (5.78)	SARS-CoV Bat coronavirus BM48-31/BGR/2008
SARS-CoV Frankfurt-1 (Cq 23)	Untargeted	6,936,399	436 (0.006)	Bovine coronavirus Porcine epidemic diarrhea virus
MERS-CoV EMC/2012 (Cq 22)	Untargeted	8,201,535	8,748 (0.1)	Bat coronavirus BM48-31/BGR/2008

[a] Enrichment by capture probes targeting vertebrate viruses designed in 2015

Building contigs



Blast results

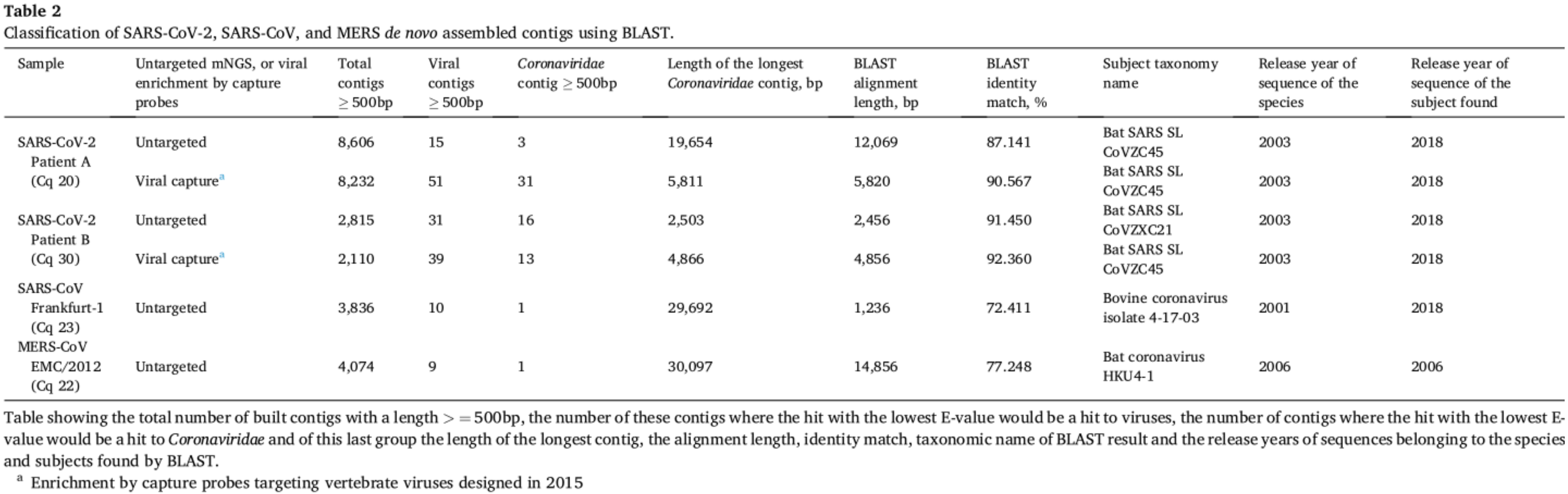


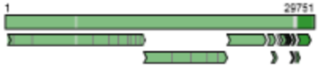
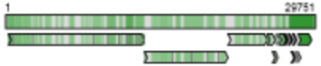
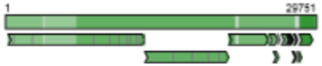
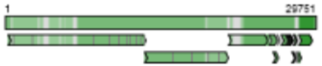
Table 2
Classification of SARS-CoV-2, SARS-CoV, and MERS *de novo* assembled contigs using BLAST.

Sample	Untargeted mNGS, or viral enrichment by capture probes	Total contigs ≥ 500bp	Viral contigs ≥ 500bp	*Coronaviridae* contig ≥ 500bp	Length of the longest *Coronaviridae* contig, bp	BLAST alignment length, bp	BLAST identity match, %	Subject taxonomy name	Release year of sequence of the species	Release year of sequence of the subject found
SARS-CoV-2 Patient A (Cq 20)	Untargeted	8,606	15	3	19,654	12,069	87.141	Bat SARS SL CoVZC45	2003	2018
	Viral capture[a]	8,232	51	31	5,811	5,820	90.567	Bat SARS SL CoVZC45	2003	2018
SARS-CoV-2 Patient B (Cq 30)	Untargeted	2,815	31	16	2,503	2,456	91.450	Bat SARS SL CoVZXC21	2003	2018
	Viral capture[a]	2,110	39	13	4,866	4,856	92.360	Bat SARS SL CoVZC45	2003	2018
SARS-CoV Frankfurt-1 (Cq 23)	Untargeted	3,836	10	1	29,692	1,236	72.411	Bovine coronavirus isolate 4-17-03	2001	2018
MERS-CoV EMC/2012 (Cq 22)	Untargeted	4,074	9	1	30,097	14,856	77.248	Bat coronavirus HKU4-1	2006	2006

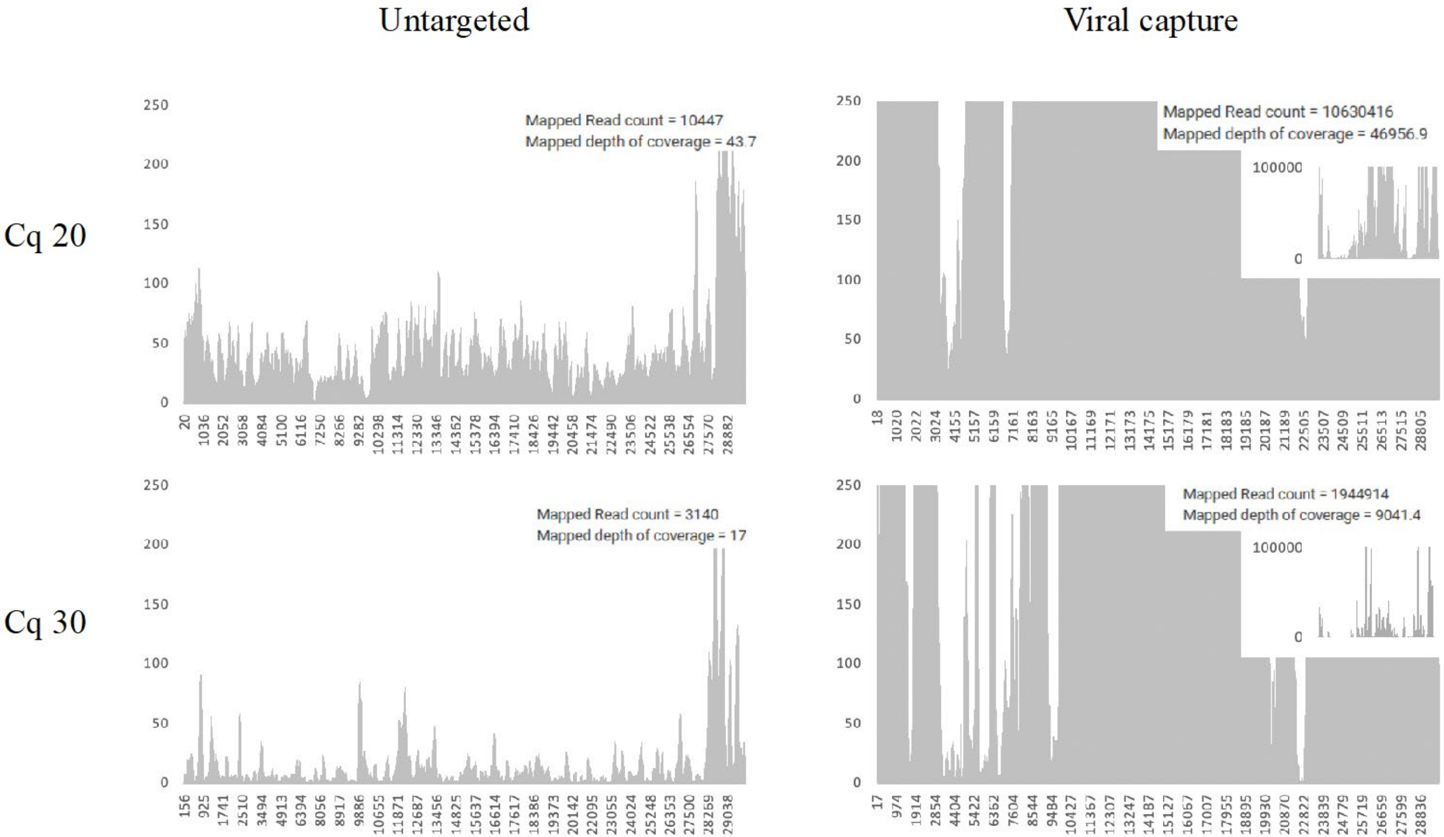
Table showing the total number of built contigs with a length > = 500bp, the number of these contigs where the hit with the lowest E-value would be a hit to viruses, the number of contigs where the hit with the lowest E-value would be a hit to *Coronaviridae* and of this last group the length of the longest contig, the alignment length, identity match, taxonomic name of BLAST result and the release years of sequences belonging to the species and subjects found by BLAST.

[a] Enrichment by capture probes targeting vertebrate viruses designed in 2015

Genome Detective Results

	Number of Contigs	Number of Reads	SARS-CoV Genome Coverage, %	Depth of Coverage	Identity, %		SARS-CoV Genome Alignment
					NT	AA	
A) Untargeted Patient A (Cq 20)	3	10,426	98.4	43.7	79.6	83.2	
Patient B (Cq 30)	36	3,126	74.2	17	80.7	84.5	
B) Captured Patient A (Cq 20)	5	10,601,614	97.1	46,956.9	80.2	83.9	
Patient B (Cq 30)	12	1,942,472	91.3	9,041.4	80.9	84.9	

Untargeted versus captured



Discussion

High and low loads of SARS-CoV-2, SARS-CoV, and MERS-CoV in clinical samples could be detected using our validation model for corona virus discovery:

- > 436 reads are classified the closest relatives of these viruses available at that time
- Clinical metagenomics protocol gave enough reads for contig assembly
- Contigs up to 14,856bp length aligned to the closest relatives of these viruses
- Low 72-92% identity of these consensus genomes with genomes of closely related ones indicated a novel coronavirus

Whole genome of SARS-CoV covered 91-97% using old capture probes

Important: nucleotide identity of over 72% to closest known relative and conclusions cannot be extended to novel viruses which are less closely related

Diagnostic implementation may contribute to increased vigilance for emerging viruses

ACKNOWLEDGEMENT

LUMC:

- Jutte de Vries
- Igor Sidorov
- Louis Kroes
- Eric Snijder
- Jeroen Laros
- Eric Claas
- Jessika Zevenhoven-Dobbe
- Margriet Kraakman
- Vreeswijk
- Lopje Höcker
- Sam Nooij
- Michel Villerius
- Leon Mei

GenomeScan:

- David van der Meer

Students

- Joost van Harinxma thoe Slooten
- Alhena Reyes

Genome Detective

- Koen Deforche
- Wim Dumon