Comparative ORF and whole genome sequencing analysis of the porcine reproductive and respiratory syndrome virus (PRRSV) in routine samples reveal a recombinant virus strain

Supplementary Document S3: *De novo* assembly using different software tools

First, we mapped all reads against *sus scrofa*. Then, the remaining unused reads were mapped against the 775 PRRSV strains (supplementary table 2). The sequence with the highest sequence coverage was set as reference sequence and the mapping was repeated against this one sequence (KT033457 for all four sequences), leading to a consensus sequence (Zhang et al., 2017).

The reads, finally mapped to the reference sequence, were also used for a following *de novo* assembly with SPAdes 3.13.0 (Bankevich et al., 2012) and with Velvet (Zerbino and Birney, 2017) within geneious. The *de novo* assembly with SPAdes (Bankevich et al., 2012) gave no results, the software stopped in all attempts with an error. Velvet 1.2.10 (K-mer length 33, min. contig length 66 bp) was partially able to assemble the reads but in lower quality. These contigs had median lengths of 80 to 186 bp. But an alignment of the *de novo* assembled contigs with the reference sequence KT003457 resulted in a coverage of only 52.8% - 83.9% and a pairwise nucleotide identity of 81% - 88.6%.

Table 1: Results of *de novo* assembly using Velvet 1.2.10
* Sample number: each sample was sequenced twice, with and without DNase
° Number of reads: reads used for the final mapping against the reference sequence KT033457
% of Ref. sequence covered: determined by an alignment of the *de novo* assembled contigs with the reference sequence KT003457.
+ Pairwise nucleotide identity: between the *de novo* assembled sequence and the reference sequence KT003457.
w/o DNAse: without DNase digestion.
w DNase: with DNase digestion.

Sample number *	Number of reads °	Number of contigs	Max. length of contigs	Median length of contigs	% of Ref. sequence covered[#]	Pairwise identity[+]
Pool 7 w/o DNase	47037	115	412 bp	80 bp	52.8 %	81%
Pool 7 w DNase	55200	146	386 bp	84 bp	55.1 %	88.8 %
Pool 8 w/o DNAse	5778	43	1266 bp	186 bp	83.9 %	88.6 %
Pool 8 w DNase	2573	Due to the low number of reads, an assembly was not tried.				

In a second attempt we mapped all reads against *sus scrofa* and used the remaining reads for a *de novo* assembly with SPAdes 3.13.0, Velvet 1.2.10 and with Tadpole, a new BBT tool, available in geneious (http://seqanswers.com/forums/showthread.php?t=61445). The *de novo* assembled contigs were aligned with the consensus sequence received by the first mapping against KT003457 (Pool 7 NCBI accession number MT857222, Pool 8 NCBI accession number MT857223).

The *de novo* assembly with SPAdes (Bankevich et al., 2012) gave again no results, the software stopped in all attempts with an error. Velvet 1.2.10 was partially able to assemble the reads. Comparable to the results of the first analysis, the coverage to the reference sequences were only around 50% and the pairwise nucleotide identity only around 75% (data not shown).

The results of the Tadpole assembly showed contigs of a median length of 242 bp to 261 bp and a higher coverage to the reference sequence (36.1% and 70.1% for Pool 7; 58.6% and 90% for Pool 8) with a very high pairwise nucleotide identity (97.6% - 100%).

We merged the two *de novo* assembled contigs originating from one sample and aligned these to the consensus sequences MT857222 and MT857223 (obtained by mapping) respectively. For Pool 7, 51 of 1305 contigs could be used for the assembly and revealed a coverage to the reference sequence MT857222 of 75.4% with 99.7% pairwise nucleotide identity. For Pool 8, 49 of 698 contigs were used, resulting in 97.2% coverage of reference sequence MT857223 and 98.4% pairwise nucleotide identity.

Table 2: Results of *de novo* assembly using Tadpole
* Sample number: each sample was sequenced twice, with and without DNase
° Number of reads: remaining reads after the mapping against *sus scrofa*
% of Ref. sequence covered: determined by an alignment of the *de novo* assembled contigs with the reference sequence.
#1 reference sequence MT857222
#2 reference sequence MT857223
+ Pairwise nucleotide identity: between the *de novo* assembled contigs and the corresponding reference sequence.
w/o DNAse: without DNase digestion.
w DNase: with DNase digestion.

Sample number *	Number of reads °	Number of contigs	Max. length of contigs	Median length of contigs	% of Ref. sequence covered[#]	Pairwise identity[+]
Pool 7 w/o DNase	130892	515	5388 bp	251 bp	70.1% #1	100%
Pool 7 w DNase	298852	790	5388 bp	256 bp	36.1% #1	98.8%
Pool 8 w/o DNAse	32636	132	3902 bp	242 bp	90% #2	97.6%
Pool 8 w DNase	190242	566	5416 bp	261 bp	58.6% #2	99.4%

The raw data of the NGS are uploaded into the NCBI SRA database under the following link https://www.ncbi.nlm.nih.gov/sra/PRJNA661439 **with the BioProjekt number PRJNA661439.**

Conclusion:

The results of the Tadpole based assembly do not show a completely continuous sequence, but show an almost perfect match in the large area covered by it. In combination with the fact, that the sequences of the supposed recombination site in the ORF5 were confirmed to 100% by an independent and alternative detection method (sanger Sequencing), we assume that Pool 7 is a recombinant PRRSV strain, origination from a MLV and a field virus.

Literature:

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012): SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 19:455-77

Zerbina DR and Birney E (2017): Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18:821-829.

Tadpole, a new BBT tool, is an extremely fast kmer-based assembler
http://seqanswers.com/forums/showthread.php?t=61445