High-throughput sequencing for discovery & genome assembly of plant viruses at the DSMZ Plant Virus Department

Dennis Knierim, Paolo Margaria, Wulf Menzel & Stephan Winter
Leibniz-Institut DSMZ-Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH, Braunschweig, Germany

- High-throughput sequencing (HTS) for virus discovery and virus genome assembly is a powerful tool to identify the genomes of unknown plant viruses and to compare virus genome variations. At the DSMZ Plant Virus Department, a workflow has been established and is extensively used to analyse RNA and DNA sequences from plant viruses.

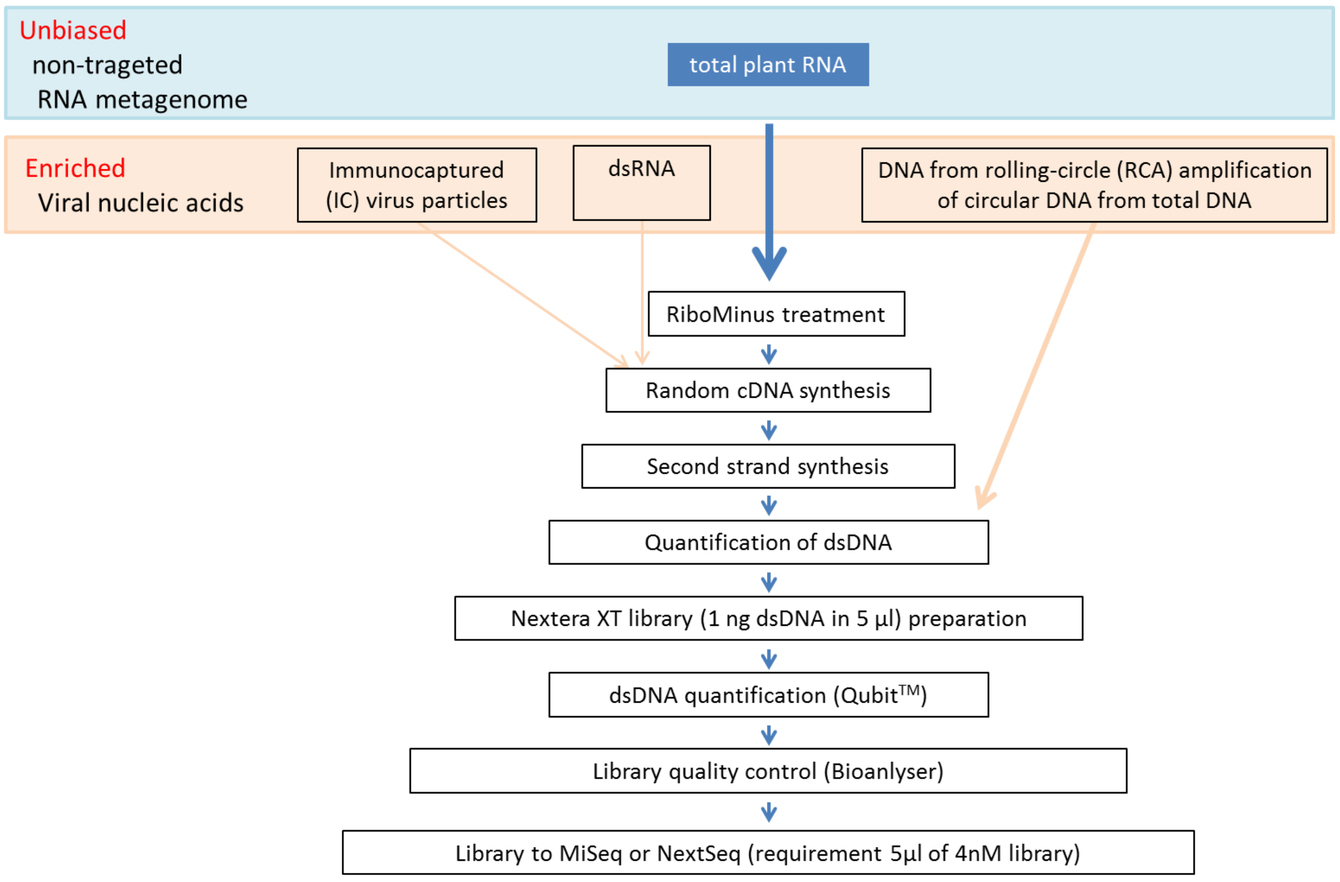

Fig. 1: Illumina library pipeline for plant virus discovery established at DSMZ.
Nextera libraries are either subjected to MiSeq (2x301) or NextSeq (2x101)

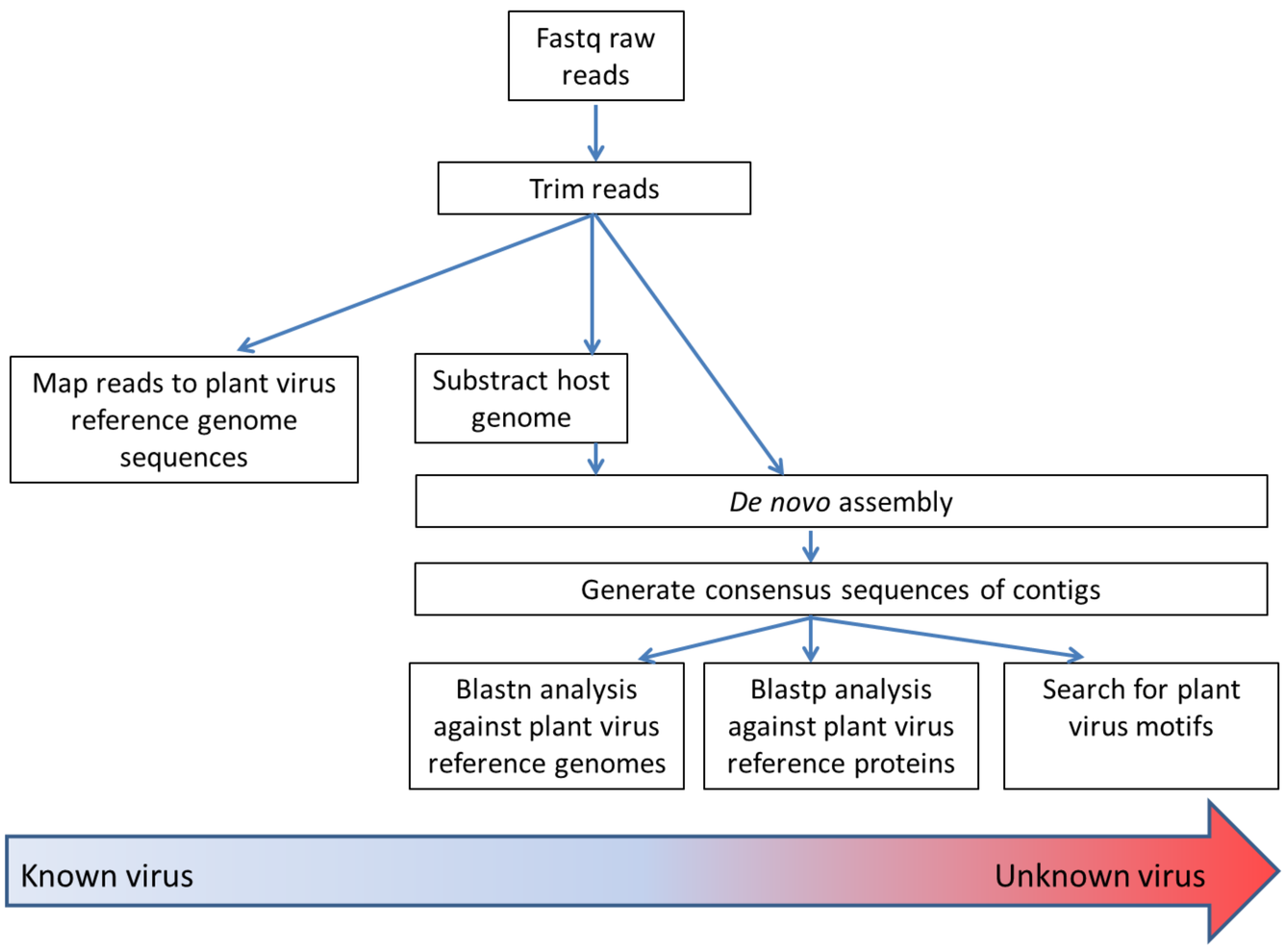

Fig. 2: Pipeline for virus sequence identification (mapping) and *de novo* assembly

Table 1: High-throughput sequencing of samples from symptomatic plants with unknown or known virus infections. Highest coverage is reached with enrichment methods

Target preparation	Infected plant	# reads	Virus/satellite identified	% reads mapped to reference
rRNA-depleted total RNA	African Eggplant	6,365,294	Polerovirus	0.08%
	Saffron	2,750,234	Beet western yellows virus	0.13%
	N. benthamiana	1,569,248	Cassava brown streak virus	2.5%
	N. occidentalis	1,007,888	Nucleorhabdovirus	3.7%
	Euphorbia sp.	2,753,264	Euphorbia ring spot virus	5.8%
IC virus particles	*D. stramonium*	1,801,708	Potato yellowing virus RNA-1	0.27%
			Potato yellowing virus RNA-2	0.23%
			Potato yellowing virus RNA-3	0.36%
	Spinach	1,474,548	Beet oak leaf virus RNA-1	3.01%
			Beet oak leaf virus RNA-2	4.64%
	Cucumber	2,550,560	Cucumber green mild mosaic virus	7.86%
	Watermelon	1,331,096	Tobamovirus	19.74%
	Potato	2,181,248	Potato leaf roll virus	21.3%
dsRNA	Cassava	2,147,068	Ampelovirus	69.76%
RCA products	Tomato	1,660,078	Tomato leaf curl Vietnam virus	46.42%
	Okra	139,982	defective Cotton leaf curl Gezira beta satellite	25.4%
			Okra leaf curl virus DNA-1	9.7%
			Cotton leaf curl Gezira alpha satellite1	7%
			Cotton leaf curl Gezira alpha satellite2	4.4%
			Okra leaf curl virus	4.4%

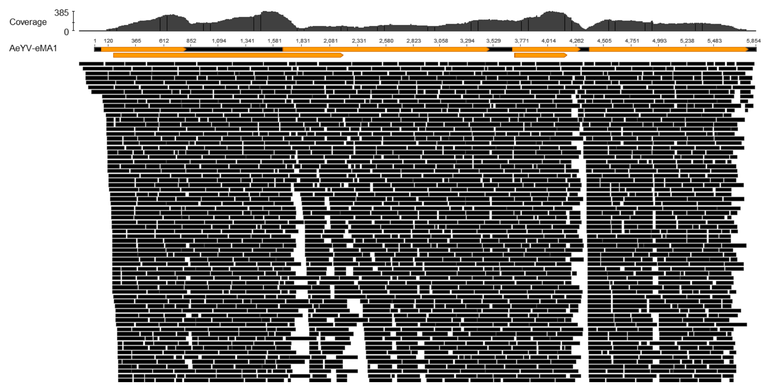

Fig. 3: Coverage of a polerovirus genome from a library of rRNA-depleted total RNA and paired end (2x301) MiSeq sequencing

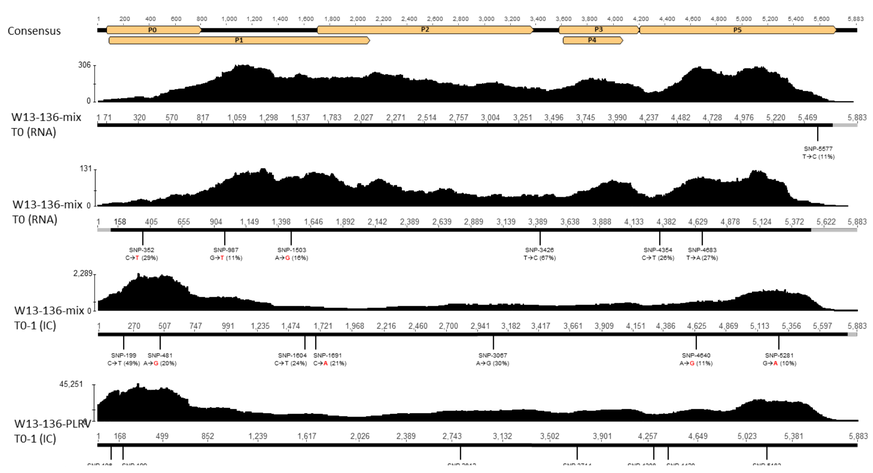

Fig. 4: Sequencing of RNA preparations of immunocaptured *Potato leaf roll virus* to identify sequence variations (SNPs) in the viral genome following sequential inoculations

- Target preparation is key to the analysis of high-throughput sequencing data for virus discovery, *de novo* assembly of virus genomes and identification of sequence variations in virus genomes. In general an unbiased method - total RNA preparation - generates much less virus specific reads, however, it covers sequences from both RNA and DNA viruses;
- An increase in the depth of the reads (>1 mio.) in general was sufficient to cover most of the virus genomes so far analysed;
- Enrichment methods are powerful – for example, RCA for geminiviruses; however they depend on the virus present, as many plant/virus infections do not result in abundant dsRNA;
- Illumina MiSeq (2x301) generates long reads with significantly facilitate *de novo* assembly of viral genomes. Only the extreme 5’ and 3’ ends (20-30 nt) are in most cases not covered.

At the DSMZ Plant Virus Department, the assessment of a plant sample with unknown virus status is performed by MiSeq analysis of total RNA providing an unbiased view on the presence of viral pathogens from all families and genera.