Text S1: General pipeline used to create semi-artificial and completely artificial datasets

Semi-artificial datasets

Analysis of the real datasets with Geneious

In this study, 8 real datasets were used. Some of them were spiked with artificial data in order to create semi-artificial datasets. These 8 real datasets are all RNA-seq datasets from virus/viroid-infected plants obtained using Illumina technology (either HiSeq, MiSeq or NextSeq sequencer). They have been chosen in order to cover as much as possible host plant diversity (fruit trees, vegetables and biological indicator plants), pathogen diversity (RNA and DNA viruses, viroids) and sequencing options (reads length ranging from 50 to 301 bp between each dataset, number of reads per dataset from 65,177 to 49,052,832 reads, and single-end or paired-end reads). All real datasets are therefore composed of plant host and virus/viroid reads. A more detailed description of these real datasets is available in Table S1.

Before the creation of the artificial reads, each real dataset was analysed in order to check that we were able to detect every viruses/viroids and to assess their presumed frequencies. Note that for each dataset, the presence of the viruses/viroids has also been confirmed by PCR and/or ELISA. The datasets were analysed in Geneious Prime 2019.2.1 (Biomatters, New Zealand). The main steps performed were:

- Pairing of the reads for paired-end datasets. For each paired-end dataset, all the reads have always been paired (i.e. no unpaired reads have been used).
- Merging/assembling of the paired reads to obtain longer reads for paired-end datasets
- Trimming of the merged and unmerged reads (BBDuk, minimum quality = Q25, minimum length = 20 bp).
- Removing duplicates of the trimmed, merged and unmerged reads. We obtain the final "filtered reads".
- De novo assembly of the filtered reads using SPAdes for the long reads (301 or 150 bp) or Velvet for the short reads (50 or 75 bp).
- Blasting of the contigs against the RefSeq virus database (https://ftp.ncbi.nlm.nih.gov/refseq/release/viral/) using tblastx from the blast+ suite (https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download).
- Mapping of the filtered reads against either the top blast hit reference sequences or the contigs (if we have a complete genome, which depends on the dataset). If a read could

be mapped with equal match to two or more strains, it was randomly assigned to one of them.

- Assessment of the proportion of each virus/viroid of the dataset.

In the end, 8 real datasets were analysed. Three were already interesting and showed challenges that we wanted to test. Therefore, these 3 datasets (No. 7, 8 and 9) were been spiked with artificial viral reads or modified. The other 5 real datasets were used to create 7 semi-artificial datasets.

Creation of the semi-artificial datasets with ART and Linux tools

ART (Huang *et al.*, 2012) is a set of simulation tools allowing the generation of synthetic next-generation sequencing reads. It mimics the real sequencing process and allows the creation of FASTQ files with single or paired-end reads of different sizes.

The first challenge was to create artificial reads showing the same quality score as the reads from the real datasets. ART allows the creation of our own quality profile files. Therefore, for each dataset, quality profiles were created using the real reads. In the case of paired-end datasets, quality profiles were generated both for the reads 1 and reads 2. This was done using the "art_profiler_illumina" tool.

Then, artificial reads showing the same size and quality score as the one of the real datasets were created using the "art_illumina" tool. The following options were used:

- qprof1: the quality profile file generated for the reads 1.
- qprof2: the quality profile file generated for the reads 2.
- p: for a paired-end read simulation (only for the paired-end datasets).
- l: the read length (bp), which is the same as the real dataset read length.
- f: the expected fold coverage. This number varies between the datasets and the strains. In the GitLab page (https://gitlab.com/ilvo/VIROMOCKchallenge/-/tree/master/Datasets), a table presenting the composition of each dataset is available. This number is provided in the "Expected average number of reads per position" column.

Figure S1 shows the results of a FastQC analysis on both the real and artificial reads generated for dataset 1. The quality scores are extremely similar between both type of reads, and the same kind of results have been obtained for the other semi-artificial datasets.

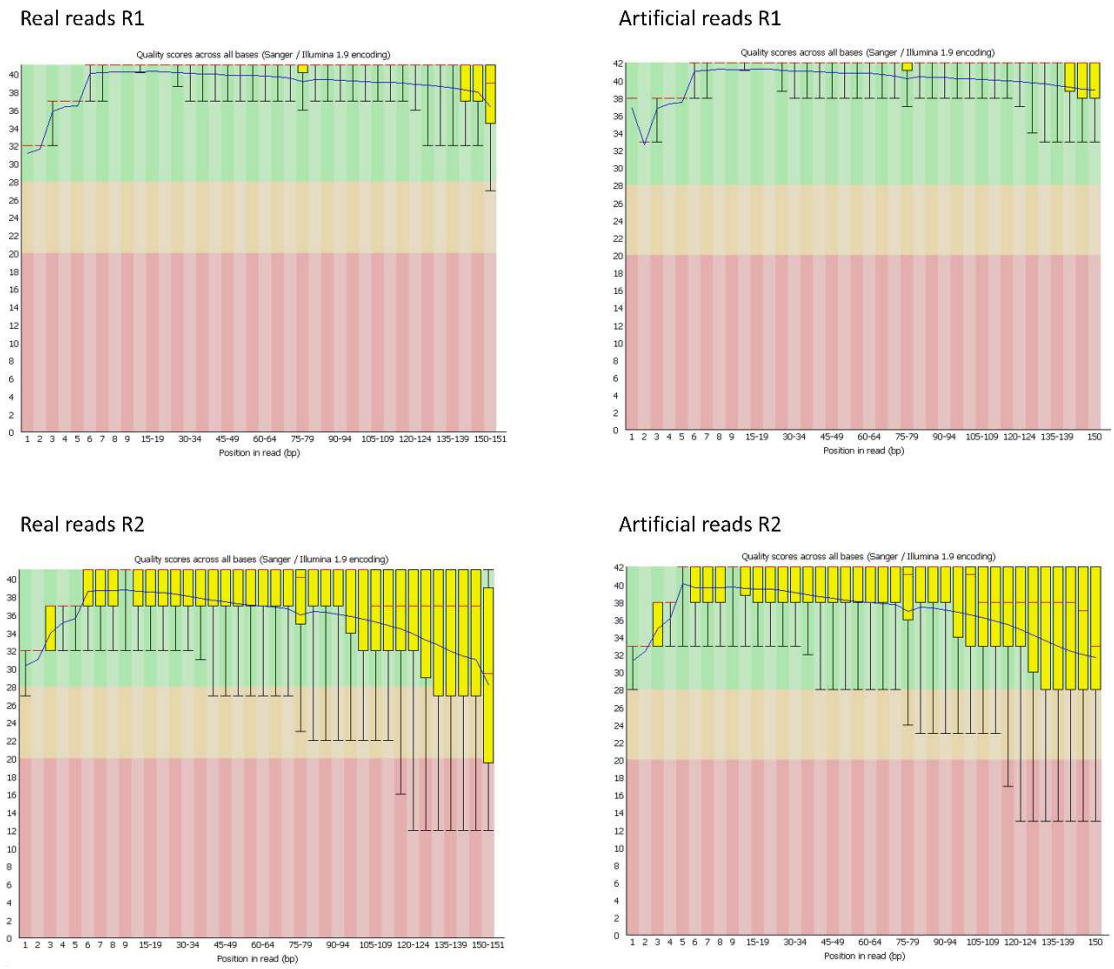

Figure S1: Quality scores of real and artificial reads of the paired-end dataset 1. The artificial reads correspond to the citrus tristeza virus (CTV) JQ911663 strain.

The second challenge was to merge artificial and real reads in the same file and with similar headers. Indeed, ART creates its own header for each read and does not allow to change it. The different steps to overcome this issue are summarized in Figure S2. These steps were performed using command-line in Linux and are the following:

- Both artificial and real reads were merged in the same file.
- The old headers were removed and new headers were added. These new headers are all the same, they look like a “real” header, but they have a number in the end. If the dataset contains x reads, x different numbers have been generated and put randomly at the end of the headers. Here is an example of headers for a hypothetical paired-end dataset with 3 reads, the number added being in red:

@E00526:48:HT7CTCXY:3:1101:16295:993:1 1:N:0:NGAGCTAC+NTAGCCTT Read 1
@E00526:48:HT7CTCXY:3:1101:16295:993:1 2:N:0:NGAGCTAC+NTAGCCTT Read 2

@E00526:48:HT7CTCXY:3:1101:16295:993:3 1:N:0:NGAGCTAC+NTAGCCTT Read 1
@E00526:48:HT7CTCXY:3:1101:16295:993:3 2:N:0:NGAGCTAC+NTAGCCTT Read 2

@E00526:48:HT7CTCXY:3:1101:16295:993:2 1:N:0:NGAGCTAC+NTAGCCTT Read 1
@E00526:48:HT7CTCXY:3:1101:16295:993:2 2:N:0:NGAGCTAC+NTAGCCTT Read 2

- The reads were sorted according to number of the header. It allowed for the mixing of the real and artificial reads.

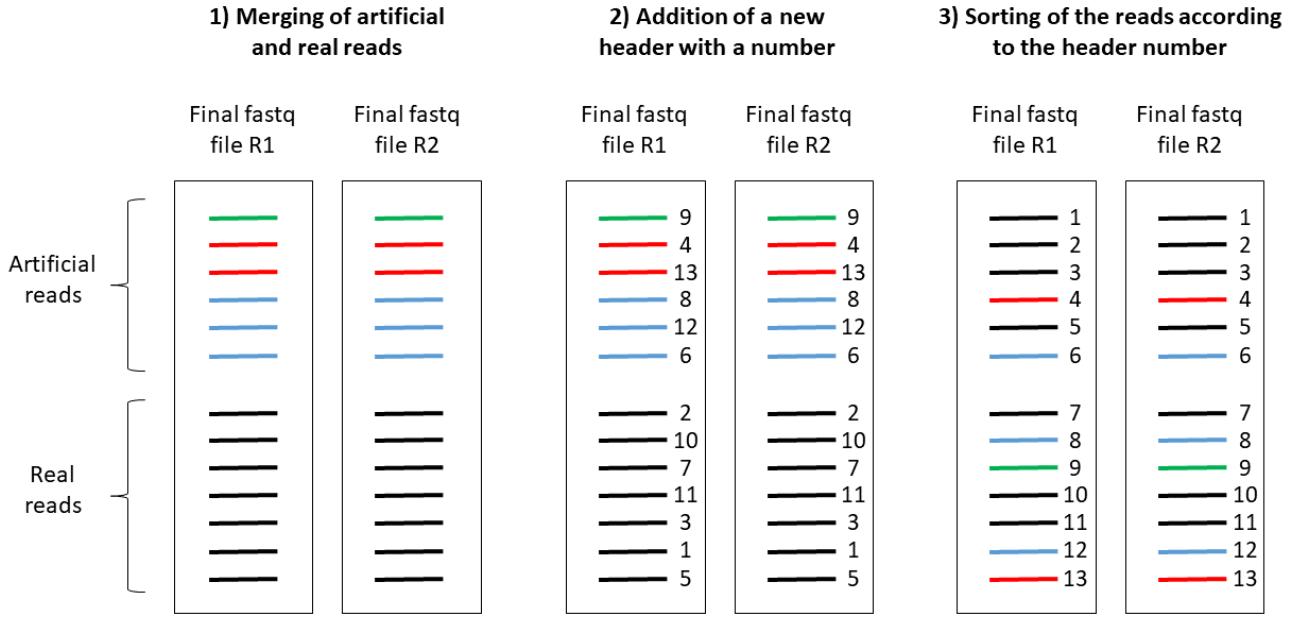

Figure S2: Final steps of the semi-artificial datasets creation.

Completely artificial datasets

Eight artificial datasets were generated (Datasets 11-18). Each artificial dataset consists of a mix of several strains from the same viral species showing different frequencies. The virus species were selected to be as divergent as possible. Therefore, the selected viruses have (i) a DNA or RNA genome, (ii) a single or double-stranded genome, (iii) a linear, circular and/or segmented genome, and (iv) show a genome length ranging from 2.8 to 17.1 kb. The virus species and strains selected are presented in the GitLab page of each dataset (https://gitlab.com/ilvo/VIROMOCKchallenge/-/tree/master/Datasets).

For each viral species, all the complete genome sequences publicly available in April 2019 were retrieved from NCBI GenBank. Then, a multiple alignment of all the sequences belonging to the same species has been performed using the "Geneious Alignement" tool from Geneious Prime (Biomatters, New Zealand) with default parameters. The obtained pairwise identity matrices have been used to select between 3 and 6 isolates per species. The isolates were selected to be as divergent as possible from each other. The average percentage of identity among isolates ranges between 62.8 (for banana streak virus) and 89.7% (for eggplant mottled dwarf virus).

For each isolate, artificial viral reads of 150 bp have been synthesized using the ART software (Huang *et al.*, 2012) from NCBI reference genomes and no single nucleotide polymorphisms (SNPs) have been added. The artificial reads were created using the "art_illumina" tool. The following options have been used:

- p: for a paired-end read simulation.
- l: 150 bp.
- f: the expected fold coverage. This number varies between the datasets and the isolates. In the GitLab page, a table presenting the composition of each dataset is available. This number is provided in the "Expected average number of reads per position" column.
- ss: HiSeq2000.

References

Huang, W., Li, L., Myers, J.R. and Marth, G.T. (2012) ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594.